Hierarchical Space-time Model Enabling Efficient Search for Human Actions

Huazhong Ning, Tony X. Han, Dirk B. Walther, Ming Liu, Thomas S. Huang, *Life Fellow, IEEE*

Abstract—We propose a five-layer hierarchical space-time model (HSTM) for representing and searching human actions in videos. From a feature point of view, both invariance and selectivity are desirable characteristics, which seem to contradict each other. To make these characteristics coexist, we introduce a coarse-to-fine search and verification scheme for action searching, based on the HSTM model. Because going through layers of the hierarchy corresponds to progressively turning the knob between invariance and selectivity, this strategy enables search for human actions ranging from rapid movements of sports to subtle motions of facial expressions. The introduction of the HIstogram of Gabor Orientations (HIGO) feature makes the searching for actions go smoothly across the hierarchical layers of the HSTM model. The efficient matching is achieved by applying integral histograms to compute the features in the top two layers. The HSTM model was tested on three selected challenging video sequences and on the KTH human action database. And it achieved improvement over other state-of-the-art algorithms. These promising results validate that the HSTM model is both selective and robust for searching human actions.

Index Terms—Hierarchical space-time model, HSTM, HIGO, action search, action recognition.

I. INTRODUCTION

Searching for human actions in a large video database or on the internet has wide-spread applications in surveillance, sports video analysis, and content-based video retrieval. This paper attempts to attack the problem of searching a human action of interest (represented by a short query video) in a large video database (called reference videos). The system queries an action against the database, and returns candidate videos containing this action and the locations of occurrences of the queried action. An intuitive idea for solving this problem is to "correlate" the short query video against the reference videos; the video locations with high behavioral similarity are selected as the matched positions. However, measuring similarity of natural human actions in video clips is very challenging due to the following reasons. One difficulty is that the same type of action, performed by two different people or by the same person at different times, is subject to large appearance variations. Another challenge is that the types of human actions vary a lot, from rapid movements of sports to subtle motions of facial expressions. Besides, the search efficiency is a critical concern.

H. Ning, M. Liu, and T.S. Huang are with Electrical and Computer Engineering Department and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: {hning2, mingliu1, huang}@ifp.uiuc.edu)

T.X. Han is with Electrical and Computer Engineering Department, University of Missouri-Columbia, Columbia, Missouri, 65211, USA (e-mail: HanTX@missouri.edu)

D.B. Walther is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: walther@uiuc.edu)

While searching for human actions in videos is rather challenging for computers, it is actually quite easy for humans. Noticing that this is also the case for object recognition in images, the answer to one question becomes essential. That is: What kind of human cognition concept or model can we borrow and leverage for computer vision?

Many scientists in neuroscience advocate a hierarchical model of increasingly sophisticated representations for processing in visual cortex [11], [21], [28], [34]. After careful study of the activation patterns of simple and complex cells in cat striate cortex, Hubel and Wiesel [14], [15] proposed a model of feature detectors in the visual cortex, starting from "simple cells" to "complex cells" and further to "hyper-complex cells". In this model, higher-level features respond to patterns of activation in lower-level cells with neighboring receptive fields, and propagating activation upwards to still higher-level cells. In accordance with the structure of the nervous system, researchers have proposed hierarchial quantitative models including Neocognitron [11], Convolutional Neural Networks (CNN) [21], and HMAX (Hierarchical Model and X) [28], for object recognition in static images.

Inspired by these successful quantitative models for object recognition in static images, we propose a five-layer hierarchical space-time model (HSTM) for searching human actions in videos. The bottom layer of this hierarchical model builds upon filter responses of 3D Gabor functions. The higher layers use the features of the HIstogram of Gabor Orientations (HIGO) or global histograms. From a feature point of view, invariance and selectivity are desirable characteristics, which seem to contradict each other. By invariance, we mean the ability of the algorithm to tolerate the changes in the human body shape, clothing, spatial-temporal scaling, background, view angle, *etc.*, given it can still recognize the correct motion. Selectivity refers to the algorithm's ability to distinguish subtle differences in motion patterns such as facial expressions or the distinction between jogging and running. To make invariance and selectivity coexist, we introduce a coarse-to-fine search and verification scheme in the HSTM model for action matching. Fig. 1 illustrates the framework of the HSTM model. We show only two layers for simplicity. Moving from the higher layers to the lower layers in the hierarchy, the feature selectivity of layers increases while their invariance to translation and scale decreases. The searching procedure starts from the highest layer, where the query video is correlated against the reference video at all locations in (x, y, t) space. The candidate locations are passed to the lower layers for further verification. Verification at lower layers enables the discrimination of subtle actions, since the features become more selective.

The hierarchical representation of the HSTM model preserves both the invariance and selectivity of the features (at different layers). The coarse-to-fine searching scheme enables it to search various human actions, ranging from rapid sports actions to subtle facial movements. This scheme, in conjunction with integral his-
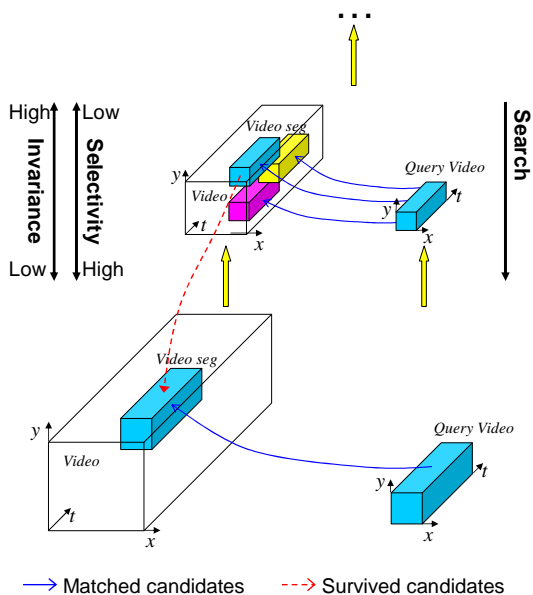
Fig. 1. The hierarchical representation of videos and the coarse-to-fine search scheme. Moving from higher to lower layers in the hierarchy, the feature selectivity of layers increases while their invariance to translation and scale decreases. The searching procedure starts from the highest layer, and the candidate locations are passed to the lower layers for verification. Here only two layers are shown. The full hierarchy has five layers.

togram [26] and HIGO features, makes the search very efficient. It takes about $1/9$ of video time to scan the entire video (see run time statistics in Section V-D). Our approach was tested on three challenging video sequences and on the KTH human action database [30]. The promising results show that the HSTM model is both selective and robust for searching human actions.

The main contributions of this paper are:

1) We propose an HSTM representation, where invariant and selective features coexist. This representation enables searching for human actions ranging from rapid sports movements to subtle facial expressions, and naturally localizes the occurrences of the queried actions in reference videos.
2) We introduce a coarse-to-fine searching scheme based on the HSTM model. This scheme improves search efficiency and realizes the desired functionality of searching human movements with a wide range of motion magnitudes.
3) We apply HIGO features and integral histograms in the high layers of the HSTM model. Both contribution 2 and 3 improve the accuracy and efficiency of searching for human actions.
4) Besides action searching, the HSTM representation is also applied to action recognition on the KTH database and is compared favorably with other state-of-the-art algorithms.

This paper is organized as follows. In the next section we focus on related work. Section III specifies our hierarchical space-time model. The coarse-to-fine search strategy, based on the HSTM model, is described in Section IV. Section V shows the experimental results. Our approach is discussed and concluded in Section VI and VII, respectively.

II. RELATED WORK

The related research falls into two sets: 1) action localization and recognition; 2) hierarchical representations for visual processing.

A. Action Localization and Recognition

The focus of the first set of literature is on action recognition, video alignment, and video matching. This literature can be classified into three categories, based on the feature representation of the video data.

The first category of approaches for action recognition relies on the analysis of spatial-temporal positions and/or shapes of human bodies in videos. Blank *et al.* [3] regard human actions as 3D shapes induced by the silhouettes while they assume fixed camera and known background appearance. Yilmaz and Shah [37] represent the posture of an action by thirteen points (landmark points) positioned on the joints of the actor. Vaswani *et al.* [35] learn the mean shape and dynamics of the trajectory change using hand-picked location data (no observation noise) and define an abnormality detection statistic. The approach by Ramanan and Forsyth [27] relies on a kinematic tracker to obtain the 2D configurations and matches them to the annotated motion capture library. Cuntoor and Chellappa [7] proposed an epitomic model for activities, using kinematics of objects within short-time intervals. These approaches are usually done with lots of human supervision, and the robustness of the algorithms highly depends on the tracking system [24].

Recently, there have been significant interests in the second category of approaches, which exploit local descriptors at interest points in videos. SIFT features [22] are probably the most widely used interest point features for image matching. The features are invariant to image scale and rotation and provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. Harris and Stephens [13] proposed an interest point detector that detects locations in a spatial image where the image values have significant variations in both directions. Based on the idea of the Harris interest point operators, Laptev and Lindeberg [20] detect local structures in space-time where the image values have significant local variations in both space and time. This representation combined with an SVM classifier was successfully applied to recognize human actions [30]. Niebles *et al.* [24] represent videos as a collection of spatial-temporal words by extracting space-time interest points and then using a probabilistic Latent Semantic Analysis (pLSA) model to categorize and localize human actions. They further propose a hierarchical model of shape and appearance for representing human actions [25]. Wong *et al.* [36] also use spatial-temporal interest points but extend the pLSA model to capture both semantic and structural information. Dollár *et al.* [9] select descriptors to characterize the cuboids of spatial-temporally windowed data surrounding a feature point. The cuboids are clustered to form a dictionary of prototypes, and then the histogram of prototypes of the video clip is supposed to reliably discriminate the human actions. An interest point describes a local region around the area of interest using representations that are distinctive enough to reliably discriminate the local region, while being invariant to a large scale of geometric perturbations and noise [17]. However, the performance of these methods may degrade due to occlusions and misdetections of the interest points [33].

The approaches in the third category attempt to measure the "behavioral similarity" using features of intensity, intensity gradient, or optical flow on the pixel level or on the space-time patch level [4]–[6], [10], [17], [33], [38]. Chomat *et al.* [5], [6] designed a set of motion energy receptive fields to sample the power spectrum of a moving texture. Then the probability density function is estimated for each class of activity by computing multi-dimensional histograms from the outputs of receptive fields. Ke *et al.* [17] use an "integral video" to efficiently calculate 3D spatial-temporal volumetric features and train cascaded classifiers to select features and recognize human actions. Boiman and Irani [4] propose to extract ensembles of local 3D patches to localize irregular action behaviors in videos. Kim *et al.* [18] introduce a framework named *Tensor Canonical Correlation Analysis* for action/gesture classification. In this framework, tensor decomposition is conducted on the video volume of raw pixel values.

These patch-based methods require no foreground/background segmentation and no motion segmentation. They tolerate appearance variance in scale, orientation, and movement to some extent. Because our approach also uses space-time patches, it shares these advantages. Our approach somewhat resembles Shechtman and Irani's approach [33] in the aspect of "correlating" a query clip to the reference videos. But we use a completely different video representation and search scheme.

B. Biologically-Inspired Hierarchical Models

Another set of literature is related to our work in the aspect of hierarchical representation for visual processing. Three typical biologically-inspired models are "Neocognitron" proposed by Fukushima [11], Convolutional Neural Networks (CNN) by LeCun *et al.* [21], and HMAX by Riesenhuber and Poggio [28]. They extended the model of simple to complex cells by Hubel and Wiesel [14], [15], and modeled the visual processing as a hierarchy of increasing sophisticated representations. Serre *et al.* [31], [32] extended the HMAX model by proposing a new set of scale and position-tolerant feature detectors, which agree quantitatively with the tuning properties of cells along the ventral stream of visual cortex. These features are learned from a set of natural images unrelated to any categorization task. Mutch and Lowe [23] built a model using Serre *et al.*'s approach [31], [32] by incorporating some additional biologically-motivated properties, including sparsification of features, lateral inhibition, and feature localization. These models were originally designed for object recognition in static images. To recognize movements in videos, Giese and Poggio [12] proposed a hierarchical model consisting of two pathways that are specialized for the analysis of shape and motion (optical flow) information respectively. This model was extended to large real datasets by Jhuang *et al.* [16].

Our approach shares with Jhuang *et al.*'s work [16] the basic idea of organizing motion features in a hierarchy with increasing complexity. However, the applications and the implementations of the top two layers are completely different. Jhuang *et al.* [16] learn a set of templates from the training set that is used for detecting spatial-temporal invariant features. This approach is originally proposed for action recognition in videos. Our application, on the other hand, is searching human actions (represented by short clips) in large video databases. The system queries an action against the database and returns candidate videos containing this action and the locations of occurrences of the queried action. This application requires high efficiency of correlating two videos

and wide applicability to videos with varying motion magnitudes. We use HIGO features and a coarse-to-fine search scheme to satisfy this purpose. The integral histograms further speed up the searching process. Therefore, our approach is application-driven as well as biologically inspired. Our approach can also be applied to action recognition after minor modifications and has achieved a recognition rate on the KTH dataset slightly better than the approach in [16].

III. HIERARCHICAL SPACE-TIME MODEL (HSTM)

Our model is partially inspired by the hierarchical quantitative models, *e.g.*, "Neocognitron" [11], CNN [21], and HMAX [28]. This type of models has achieved good quantitative results for recognizing objects in cluttered natural scenes [31]. It consists of computational units of "S" layers and "C" layers that are analogies of V1 *simple cells* and *complex cells* [14], [15]. The "S" layers combine their inputs to increase object selectivity by convolving the previous layer with local filters. The "C" layers boost the invariance to a range of scales and positions by pooling spatially over limited ranges in the previous layer. Riesenhuber and Poggio [28] advocated the max-like pooling operation, which is also used in this paper but in the space-time domain. These models have exhibited significant success on object recognition in static images and stepped further to bridge the gap between computer vision and neuroscience.

Hubel and Wiesel did not specifically propose a biological model for visual perception of spatial-temporal motions. But we find that a quantitative model in the spatial-temporal domain based on the idea of V1 *simple cells* and *complex cells* [14], [15], is very effective in visual searching. Our HSTM model can be regarded as an extension of its purely spatial counterparts in [23], [28], [31], considering its general organization of S and C layers. But the implementation is quite different, especially at the top two layers. This implementation is also motivated by the specific requirement of search efficiency and 3D characteristics of space-time videos. In summary, our model is application-driven as well as biologically inspired.

A. The Five Layers of the HSTM Model

Our model consists of five layers. Following the naming conventions of [11] and [28], we refer to the top four layers as S1, C1, S2, and C2 as shown in Fig. 2. It is worthy to notice that these layers are defined in 3D space-time. The lowest layer is the original video. The S1 responses are obtained by convolving the original video with a bank of 3D Gabor filters. Pooling over limited ranges in the S1 layer through a max-operation results in the C1 layer. We adopt the histogram of Gabor orientations (HIGO) features for the S2 layer and 3D Gabor coefficient histograms for the C2 layer. Using 3D integral histograms significantly reduces the computational cost for computing the features in the S2 and C2 layers. Our S2 layer is built from histograms of Gabor orientations, which makes it unlike simple or complex cells in the brain. Nevertheless, we decided to keep the conventional name.

Video Layer. The query clip is assumed to be much shorter than the reference video. The reference video frames are down-sampled to spatial resolution 320×240 or smaller (while maintaining the aspect ratio) to save computations. The temporal resolution is preserved.

S1 Layer. S1 responses are obtained by convolving the video with a bank of 3D Gabor filters. The Gabor filters are composed
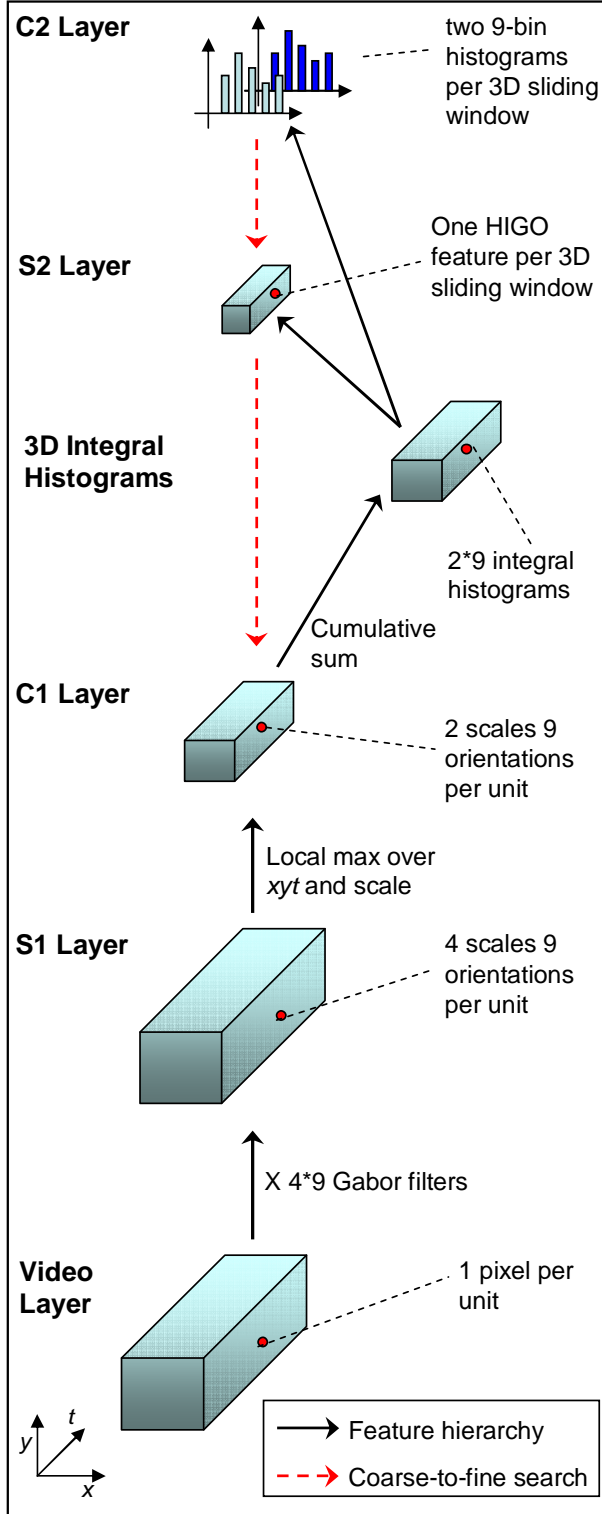
of two main components, a sinusoidal carrier and a Gaussian envelope. They exhibit many properties common in neurons in mammalian primary visual cortex, such as spatial localization, orientation selectivity, and spatial frequency characterization. The bank of 3D Gabor filters used in this paper are

$$G(x,y,t) = \exp\left[-\left(\frac{X^2}{2\sigma_x^2} + \frac{Y^2}{2\sigma_y^2} + \frac{T^2}{2\sigma_t^2}\right)\right]$$
$$\times \cos\left(\frac{2\pi}{\lambda_x}X\right)\cos\left(\frac{2\pi}{\lambda_x}Y\right) \quad (1)$$

$$\begin{pmatrix} X \\ Y \\ T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix}$$
$$\times \begin{pmatrix} \cos(\omega) & 0 & \sin(\omega) \\ 0 & 1 & 0 \\ -\sin(\omega) & 0 & \cos(\omega) \end{pmatrix}\begin{pmatrix} x \\ y \\ t \end{pmatrix}. \quad (2)$$

Here θ and ω are used to selectively rotate the filter to particular orientations in 3D space. Both θ and ω take three discrete values $-\frac{\pi}{4}$, 0, and $\frac{\pi}{4}$, so that there are 9 orientations in total. The other filter parameters (the filter size, the effective width σ, and the wavelength λ) are determined by considering the profiles of V1 parafoveal simple cells [14], [15] and the HMAX setup for 2D static images in [31]. Compared with 8 scale-bands in [31], we choose only 4 scales that are formed into 2 scale-bands (2 scales per band), to save computational cost. Hence, there are 4×9 filters. In our experiments, no significant improvement of performance was achieved with more scale-bands.

C1 Layer. The C1 computational units correspond to the complex cells that have larger receptive fields and respond to the oriented bars and edges within the receptive field. It tolerates large variances of scales and positions. To simulate complex cells, our pooling operation takes the maximum over ranges of size $8 \times 8 \times 4$ and step size $4 \times 4 \times 2$ in each scale, and then takes the maximum over the two scales in each band. After this pooling, there are 2 scale-bands (now one scale per band) and 9 orientations per unit, and the video size is down-sampled by $4 \times 4 \times 2$.

S2 Layer. During search, the reference video is scanned by sliding a 3D window over every space-time position. The similarity between each sliding window and the query video is computed. For each 3D sliding window and the query clip, the histogram of the Gabor orientations (HIGO) is extracted from the C1 layers. Our HIGO feature is similar to the histogram of gradient (HOG) feature proposed by Dalal and Triggs [8]. Extraction of the HIGO feature consists of the following steps.

1) The C1 layer of the 3D sliding window/query video is subdivided into non-overlapping cells, consisting of $4 \times 4 \times 3$ C1 units.
2) The C1 responses in each cell are accumulated for the 9 orientations respectively, forming a 9-bin histogram. Because of its summarizing characteristics, the histogram exhibits a range of scale- and position-invariance.
3) Cubes of $2 \times 2 \times 2$ cells in the 3D sliding window form blocks. Adjacent blocks have 50% overlap. Fig. 3 shows the 3D sliding window, cells, blocks, and their relationship.
4) For each block, the 9-bin histograms of its 8 cells are concatenated to form a 72-dimensional vector v.
5) We perform *L2-Hys* normalization over the 72-dimensional vector v for each block, as in [22]: first it is l_2-normalized, $v \rightarrow v/\sqrt{\|v\|_2^2 + \epsilon^2}$, followed by clipping (limiting the

Fig. 2. Overview of the five layers of the HSTM model. Both the reference video and the query clip are processed in the same way. During search, the reference video is scanned by sliding a 3D window at all positions in the C2 layer and at candidate positions in the lower layers. The query clip is matched against all sliding windows. Note that the sliding window and the query clip have the same size in each layer.
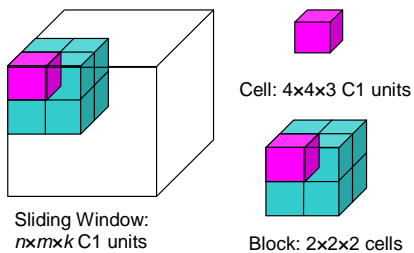
Fig. 3. Relationships between 3D sliding window, blocks, and cells. The sliding window is subdivided into non-overlapping cells. Cubes of $2 \times 2 \times 2$ cells in the 3D sliding window form blocks. And adjacent blocks have 50% overlap. Here only one block is drawn in the sliding window.

maximum values of v to be 0.1), and then it is normalized again. Here ϵ is a small constant (10^{-6} in this paper) to avoid zero denominator.

6) All the normalized vectors of the blocks in the 3D sliding window are again concatenated in a fixed order to retain the geometric information for selectivity. This concatenation forms the final HIGO feature of the window. The dimension of the HIGO feature depends on the query video size.

We cannot extract the HIGO features for the video database beforehand, because the size of the sliding window is determined by that of the query clip, which may vary. This would pose heavy computational load on the search procedure. However, an intermediate layer can be inserted between the C1 and S2 layers that computes the integral histograms over C1 responses. Computations on the S2 and C2 layers are done on this intermediate layer, instead of the original C1 responses. This saves most of the computation time, because integral histograms can be computed and saved beforehand, and building a histogram over a cuboid region needs only 7 "add/subtract" operations for the 3D case. Also note that our S2 and C2 layers are quite different from those in [16], [23], [28], [31] due to the special requirements of searching in the 3D spatial-temporal domain.

We adopt HIGO features on the S2 layer for the purpose of efficient and fast searching of human actions. The success of HIGO features is mainly due to: 1) using Gabor responses to emphasize bars and edges; 2) computing the histograms over small cells to obtain scale- and position-invariance; 3) keeping the order of cells to maintain geometric information, *i.e.*, to maintain selectivity; and 4) efficient computation with the support of integral histograms.

C2 Layer This layer is the top layer, where the search is initiated. The scanning window slides over all positions in this layer (See Section IV). Therefore, the cost for feature extraction and similarity measurement in this layer should be as low as possible. We choose the histogram of the Gabor filter orientations over the entire scanning window as the C2 feature. This histogram is computed as follows: the C1 response in the entire scanning window is accumulated for 9 orientations, respectively, and the 9 accumulations form a 9-bin histogram. Facilitated by the integral histograms ready in the intermediate layer, it takes only 7 "add/subtract" operations for the C2 computational units to compute a 9-bin histogram. The feature at the C2 layer captures the global motion information, and it does not maintain the geometric information. In other words, it is highly invariant while with low selectivity. Also this feature has very low dimensionality (2×9

dimensions) and leads to efficient similarity measurement. This exactly satisfies the requirement of quickly selecting candidates at the C2 (top) layer.

IV. COARSE-TO-FINE SEARCHING IN HSTM

In the HSTM model, the higher the layers are, the fewer dimensions the features have, and therefore, less computational cost is needed to compute their similarities. Thus it is natural to propose a coarse-to-fine search strategy to improve efficiency. We start with an exhaustive scanning from the top (C2) layer and pass the possible candidate locations to the lower layers for further verification.

Furthermore, the hierarchically represented features are increasingly more invariant, more global, and less selective, when they are examined from low to high layers. This allows another advantage of the coarse-to-fine strategy, *i.e.*, it enables the separation of human actions with a wide range of varying motion magnitudes by verifying the candidate locations through different combinations of layers.

For the C2 layer scan, we choose histogram-based features that are calculated quickly from the pre-computed integral histograms. The query video is exhaustively searched on the entire layer, *i.e.*, the similarity between the query video and the scanning window is measured at every location in the C2 layer. The locations with the highest similarities are picked up as candidates. Note that the video size at the C2 layer is very small after down-sampling at the C1 layer. Hence the exhaustive search is very fast.

Using histogram features as action representation allows for the use of divergence functions from information theory and statistics directly for similarity measurement. Two of the most prominent divergences are Kullback-Leibler (KL) divergence and χ^2-divergence [29]. We chooses the KL divergence [19] to measure the similarity between two histograms f and g:

$$\begin{aligned} KL(f||g) &= \int f(x) \log \frac{f(x)}{g(x)} dx \\ &= \int f \log f dx - \int f \log g dx \end{aligned} \quad (3)$$

To make the measure symmetric, we take the symmetric KL divergence: $d(f,g) = KL(f||g) + KL(g||f)$. Then the similarity A between the query clip and the 3D sliding window in the reference video can be computed by

$$A(h_s, h_q) = \exp\left(-\frac{d^2(h_s, h_q)}{2\sigma^2}\right) \quad (4)$$

where h_s and h_q indicate the histograms of the sliding window and query video at the C2 layer, respectively, and $d(h_s, h_q)$ is the symmetric KL divergence between the two histograms (see **C2 layer** in Section III-A). σ is the variance of the KL divergence over all of the sliding windows. We select locations with similarity measures in the top 0.5% as possible candidates. Duplicate candidates are consolidated using the neighborhood suppression algorithm from [2]. Most of the candidates are false alarms and should be rejected, because the histogram-based features lack selectivity. But true matches are selected with high probability. Usually, the real locations rank far higher than the top 0.5% in the similarity measure.

These possible candidates are passed to the lower (S2) layer for verification. The dimension of the HIGO features for a query video (or a 3D sliding window) with size $100 \times 100 \times 30$ is as high

as $7,000$. Therefore we choose the l_1 norm to compute the similarity for efficiency. In our experiments, it achieves performance similar to the l_2 norm. HIGO features are very discriminative to separate the positive from the negative candidates, with their good property of trading off between invariance and selectivity. Usually, the search can be terminated at this stage.

However, if the actions to be queried are as subtle as facial expressions instead of rapid movements, HIGO features are not selective enough to discriminate these fine motions. Therefore, the surviving candidates from the S2 layer need to be further verified at the C1 layer. Here, the simple SSD (Sum of Squared Differences) can be used to compute the similarity since only a few candidates need to be verified, and the total computational load is low. We determine whether the candidates should be further verified at the C1 layer by considering the average magnitude of motion in the query video. Assume $c_1(x, y, t|\theta, \omega)$ is the response at (x, y, t) on the C1 layer of the query video, with Gabor orientation (θ, ω) (θ, ω are defined in Eqn.2), and

$$C_1(\theta, \omega) = \int_x \int_y \int_t c_1(x, y, t|\theta, \omega)dxdydt \qquad (5)$$

is the total response in the direction of (θ, ω). When θ and ω approach zero, the Gabor orientation points to the t axis and responds to the pixels with small motion. Therefore, the ratio

$$P(\epsilon) = \frac{\int_{|\theta| < \epsilon} \int_{|\omega| < \epsilon} C_1(\theta, \omega)d\theta d\omega}{\int_\theta \int_\omega C_1(\theta, \omega)d\theta d\omega} \qquad (6)$$

reflects the motion magnitude of the query video. Here ϵ is a fixed small positive number. A small motion magnitude usually corresponds to a large ratio $P(\epsilon)$. So the candidates should be further verified at the C1 layer if $P(\epsilon)$ is larger than a threshold.

It is worthy to mention that, for the query video of subtle motions, the ratios $P(\epsilon)$'s, $i.e.$, the bins of the histogram around $\theta = \omega = 0$, are dominant but contribute little to motion discrimination. To suppress their influence, the histograms bins at the C2 and S2 layers are weighted by a prior to suppress the dominant entries around $\theta = \omega = 0$.

We never verify the candidates at the S1 layer because its features are too selective and lack invariance. Hence, the S1 response does not need to be stored for the querying, which saves storage space.

V. EXPERIMENTAL RESULTS

Our hierarchical representation and coarse-to-fine search strategy has wide applications ranging from internet video searching, video surveillance, to sports video analysis. It can be used to retrieve human actions ranging from rapid body movements (such as playing tennis) to subtle motions such as facial expressions during web conversations. We search with a short query video, which represents the human action of interest, in the reference videos, and return the locations of similar actions. This method requires no background/foreground segmentation and can tolerate a large range of scales, positions, and motion variations. We carried out extensive experiments on both selected challenging videos and public human action database ($i.e.$, the KTH database). Please view the video clips (queries, candidates at C2, S2, or C1 layers, and the similarity surfaces) at our website [1].

A. Query in Selected Videos

Fig. 4 shows the results of searching strokes in a tennis video. The short query video is a tennis stroke of 31 frames of 104×124 pixels. Fig. 4 (a) shows a few frame samples. The query video is searched in a tennis reference video (800 frames of 228×146 pixels). We build an HSTM model of both the query and the reference videos (in real applications most of the model can be computed beforehand) and start the search from the C2 layer. About 300 candidate locations with high similarities are selected. Fig. 4 (b) shows a few candidates. The frames of a candidate sliding window are marked by rectangles with the same color; different colors means different candidates, while the same color not necessarily refers to the same candidate because the colors are repeated). For illustration, only the 30% of the selected candidates with the highest similarities are marked in the C2 layer. All candidates selected at the C2 layer are passed to the S2 layer for verification. Fig. 4 (c) shows all candidates that pass this verification. All of the stroke instances in the long video are captured at this stage. Fig. 4 (d) and (e) draw the similarity surfaces corresponding to frames in (b) and (c), where *red* indicates high similarity and *blue* low similarity. The surface peaks at the C2 layer are much flatter than those at the S2 layer. The C2 layer includes all of the peaks at the S2 layer while containing many spurious peaks, which are suppressed at the S2 layer. This means that the features at the S2 layer are more selective and less invariant.

Note that in Fig. 4, Fig. 5, and 6, the "correlation" between the reference and query video is computed without zero-padding, so that the surface size is equal to the reference video size minus the query video size. Here we scale the surface width to that of the reference video frame for better illustration. Furthermore, the surface peaks are not exactly aligned with the video frames in the illustration, because of the scaling and the fact that the query clip and the sliding window are aligned at the top-left corner of the first frame but not the center. Some frames in the figures are marked as candidates while the corresponding similarity surfaces are flat. This is due to that they are not the first frames of the located actions, and the peaks fall off quickly beyond the first frames. Actually, salient peaks would be seen closely preceding those frames in the videos. Readers are referred to the videos on our website [1] for better illustration.

Fig. 5 shows the results of searching turn actions in ballet footage[1]. It contains 400 frames of 192×144 pixels. The query video is a single turn of 20 frames with resolution 96×122. Some sample frames are shown in Fig. 5 (a). Fig. 5 (b), (c), (d), and (e) show some candidates at C2 and S2 layers and their corresponding similarity surfaces, respectively. This example is challenging because: 1) the video contains fast moving parts; 2) the female dancer is wearing a skirt; 3) the variability in scale relative to the template is large. Our method detects most of the turns of the two dancers. Shechtman and Irani [33] have tested their method on this video using the same query video. Comparison shows that both approaches achieve similar performance (two missed turns).

Fig. 6 shows that our approach is capable of searching subtle motions. The set of parameters do not need to be tuned specifically for the subtle motions. The query is a video clip of a person smiling from a web conversation, containing 32 frames of

---

[1]It was performed by Birmingham Royal Ballet. The original video is available on the website: http://www.londondance.com/video/brb.wmv
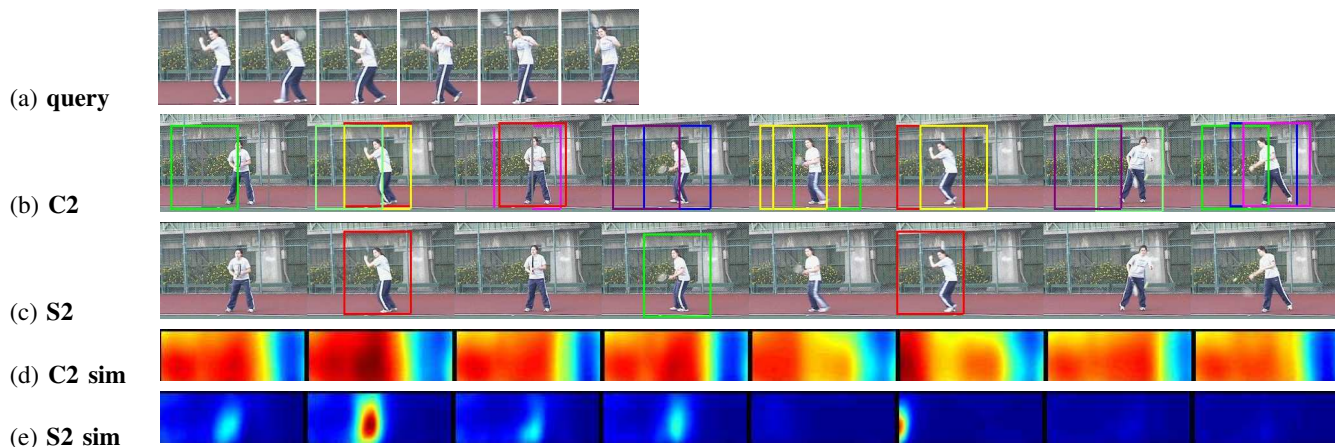
Fig. 4. **Tennis.** (a) Query video of a stroke. (b) and (c) show some candidate locations at the C2 and S2 layers, respectively. (d) and (e) are the similarity surfaces corresponding to the frames in (b) and (c), respectively. Note that the surface peaks are not exactly aligned with the video frames due to scaling, and that some marked frames have flat similarity surfaces, because they are not the first frames of the located actions. See detailed explanation in the text. See the videos on our website [1] for better illustration.
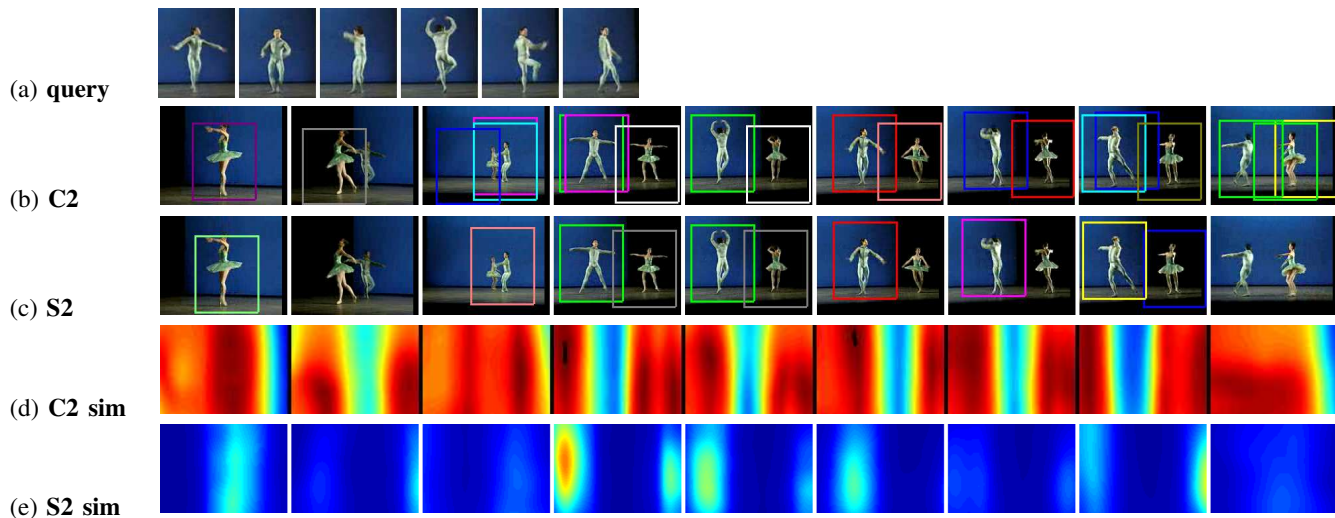

Fig. 5. **Ballet.** (a) Query video of a single turn. (b) and (c) show some candidate locations at the C2 and S2 layers, respectively. (d) and (e) are the similarity surfaces corresponding to the frames in (b) and (c), respectively. Note that the surface peaks are not exactly aligned with the video frames due to scaling, and that some marked frames have flat similarity surfaces, because they are not the first frames of the located actions. See detailed explanation in the text. See the videos on our website [1] for better illustration.

82×106 pixels. Some sample frames are shown in Fig. 6 (a). It was searched in a reference video (800 frames of 200×150 pixels) that contains large head movements and various facial expressions (happy, sad, neutral, *etc.*). Note that the faces in the query video and the reference video are taken from different subjects. Fig. 6 (b), (c), and (d) show some candidates at the C2, S2, and C1 layers, respectively. For this searching task, the candidates selected at the C2 layer need to be further verified at the C1 layer, because the ratio $P(\epsilon)$ in Eqn. 6 is greater than the predefined threshold. Although the motion of smiling is very slight, and the head movement is big, the coarse-to-fine searching is able to detect all instances of smiling. This experiment also suggests an approach to emotion recognition/classification in video sequences.

B. Query on the KTH Database

The KTH human motion database [30] is one of the largest available video sequence datasets of human actions. The video database contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping), performed several times by 25 subjects in 4 different scenarios: outdoors $s1$, outdoors with scale variation $s2$, outdoors with different clothes $s3$, and indoors $s4$. Some sample videos are shown in Fig. 7. Note that the KTH dataset contains large variations in human body shape, view angles, scales, and appearance (clothes). Our experiments show that our model can basically tolerate these variations.

We carry out experiments of searching for actions on the KTH database, *i.e.*, searching the database using a query clip and returning the candidate sequences with highest similarities. "Leave-one-out" cross validation is used to measure the performance. Each time, one sequence is selected as the query clip and is searched on all other sequences in the database (leave-one-sequence-out, LOSO) or on the sequences of the other 24 subjects (leave-one-person-out, LOPO). Instead of using the entire sequence as the
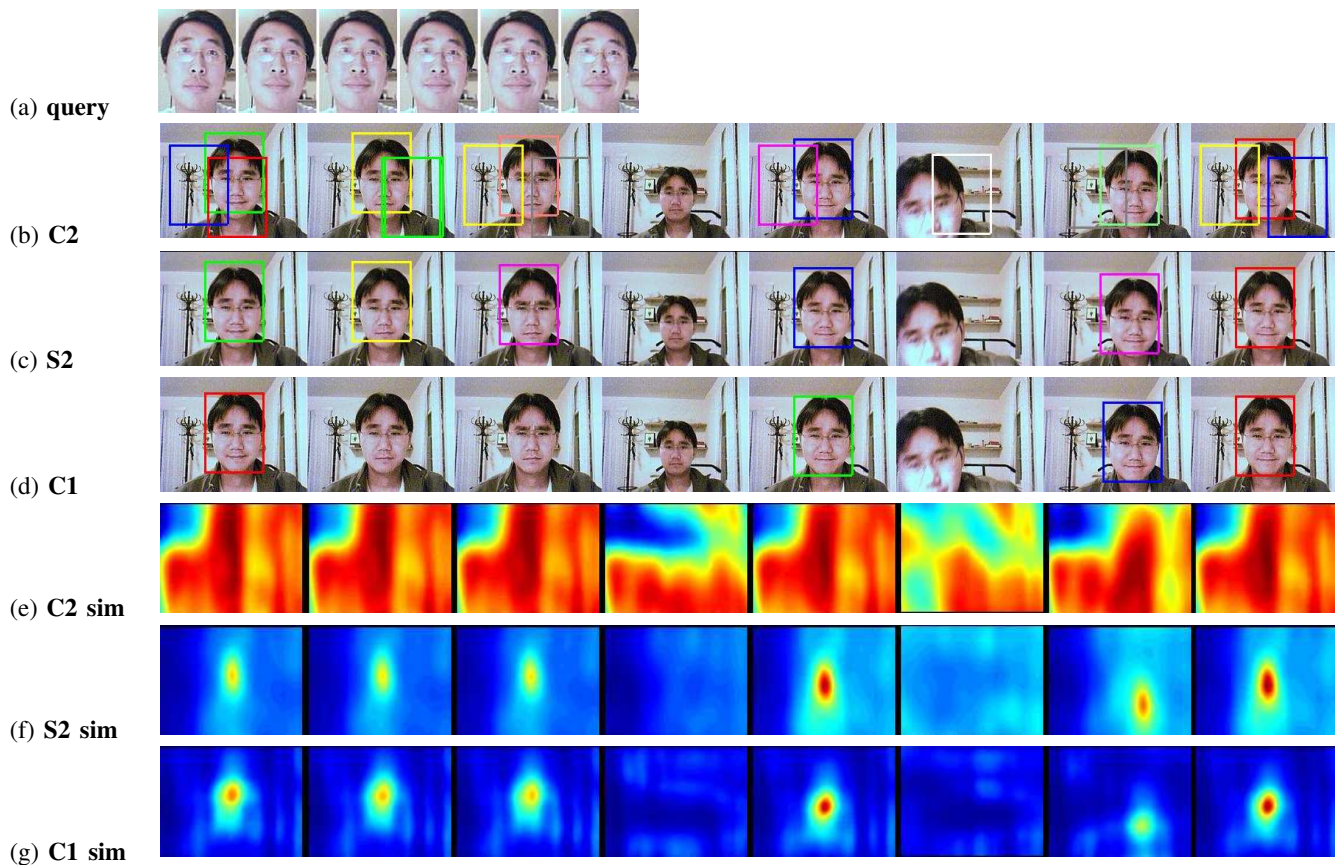
Fig. 6. **Smiling.** (a) Query video of a person smiling. (b), (c), and (d) show some candidates at the C2, S2, and C1 layers, respectively. (e), (f), and (g) are the similarity surfaces corresponding to the frames in (b), (c), and (d), respectively. Note that the surface peaks are not exactly aligned with the video frames due to scaling, and that some marked frames have flat similarity surfaces, because they are not the first frames of the located actions. See detailed explanation in the text. See the videos on our website [1] for better illustration.
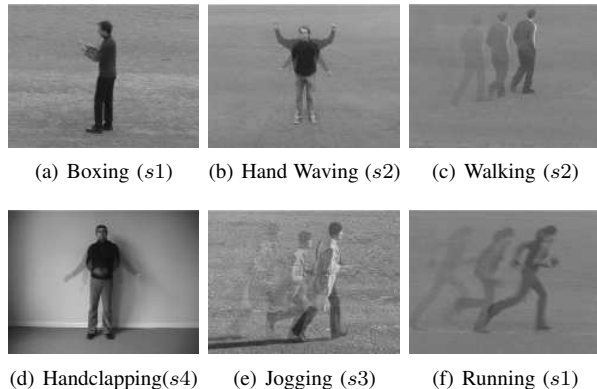

Fig. 7. Example videos from the KTH dataset. (a) Boxing ($s1$), (b) Hand Waving ($s2$), (c) Walking ($s2$), (d) Hand clapping ($s4$), (e) Jogging ($s3$), and (f) Running ($s1$).

query, we have developed a procedure that automatically crops from the original sequence a small 3D window with 1 second length that confines the human action. This 3D window is chosen as the query clip. Then the query clip is matched against the remaining videos in the database using the coarse-to-fine search scheme. On average, the search time is about $1/9$ of the video time given that the integral histograms are extracted beforehand (see Section V-D). Also note that the system returns not only the candidate video sequences, but also the locations in the sequences where the query clip is matched. These returned locations would help the users to browse the returned videos.

For each query clip, the reference videos are ordered according to the maximum values of the corresponding similarity surfaces, and the top n are returned. Fig. 8 plots the average precision and recall of searching, with n varying from 1 to 99 or 96 (For the KTH dataset, each ideal query has 99 true positive returns for LOSO and 96 for LOPO). For LOSO in Fig. 8, precision is 82.19% when $n = 10$, which is promising for searching human actions in large video dataset. An interesting observation is that the precision of LOPO is slightly higher than that by LOSO when n is small and recall is higher when n is large. The main reason is that it is difficult to discriminate jogging and running actions of the same subject, which contributes to higher errors of the LOSO setup.

We further test the robustness of our model by varying the length of the query clips and by scaling them in time. The experiments are conducted under the LOPO setup. In the first experiment, the query clips are cropped out from the original sequences in the same way as mentioned above, but their lengths are 1 second, $2/3$ seconds, and $1/3$ seconds. Fig. 9 reports precision and recall curves. Basically, the longer the query clips, the better the performance. But, when the query clips are long enough, *e.g.*, $2/3$ seconds, further lengthening the query clip makes only a slight improvement. And very short query clips,
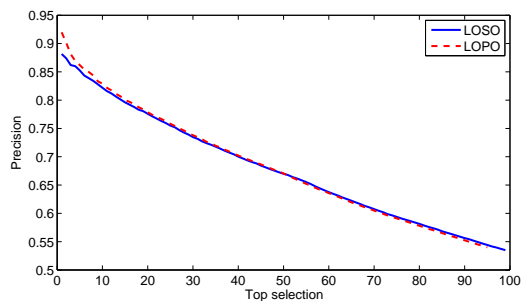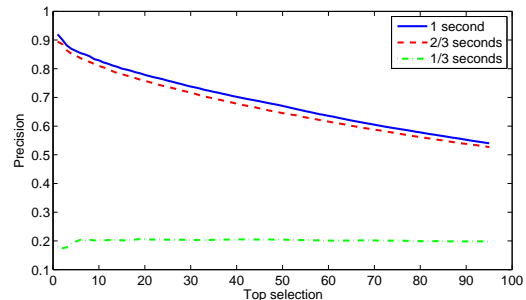
(a) Precision
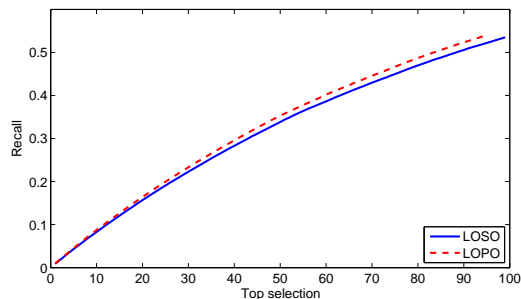
(b) Recall

Fig. 8. Precision and recall on the KTH database for both leave-one-sequence-out (LOSO) and leave-one-person-out (LOPO) setups. (a) Precision; (b) Recall.

(a) Comparison of precision

(b) Comparison of recall

Fig. 9. Comparison of precision/recall when the length of query clips is 1, 2/3, or 1/3 seconds. The experiments are conducted on the KTH database under the LOPO setup. (a) Comparison of precision; (b) Comparison of recall.

*e.g.*, 1/3 seconds, may result in very bad performance.

In the second experiment, the query clips are initially cropped with 1 second length, but then interpolated in time-axis by a scale of 0.7, 1, or 1.3. We choose these scales because, if a temporal variation of an action is beyond the interval $(0.7, 1.3)$, this variant action can basically be viewed as a different action, as walking *vs.* jogging and jogging *vs.* running. Fig. 10 gives the precision and recall under different scales. Although temporal variations might degrade the performance, the extent of degrading is very small when the variation is ranging within the interval $(0.7, 1.3)$, especially when the variation is temporal shrinkage. This demonstrates that our model is robust to temporal distortion in a range of at least $(0.7, 1.3)$. We do not conduct extra experiments to test the robustness with respect to spatial distortions, because the KTH dataset itself contains large variations of view angle and spatial scale (*e.g.*, scenarios $s2$), appearance (*e.g.*, scenarios $s3$), background (*e.g.*, scenarios $s4$), *etc.* All of above results are obtained under these variations.

C. Action Recognition on the KTH Database

Our HSTM model is originally designed for searching actions in video databases. It can also be applied to human action recognition after minor modifications. To compare with the performance of other methods [9], [16]–[18], [24], [30], [36], we conduct experiments on the KTH dataset under three experimental setups. **Setup 1** uses the same training and testing sequences as in [17], [30], *i.e.*, 8 people for training and 9 for testing. **Setup 2** performs leave-one-out cross validation similar to [9], [18], [24], [36], *i.e.*, for each run one sequence of a subject is chosen for testing and the sequences of the other 24 subjects as training set. **Setup 3** follows the experimental setup in [16] where 16 subjects are randomly drawn for training and the remaining 9 for testing. The

(a) Comparison of precision
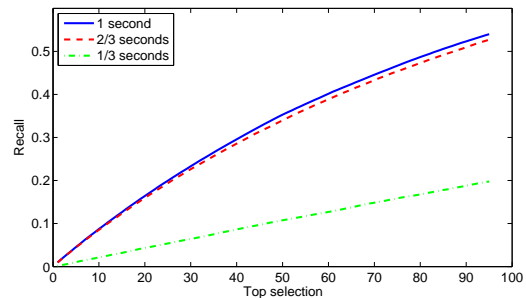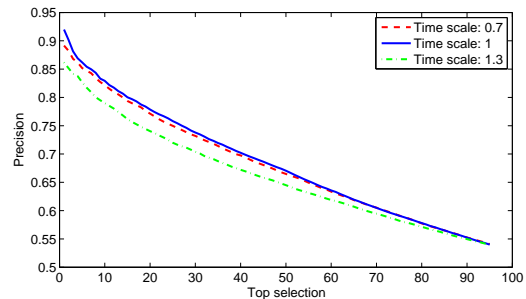
(b) Comparison of recall
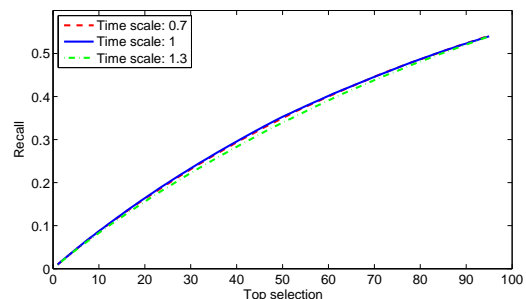
Fig. 10. Comparison of precision/recall when the query clips are interpolated in time by a scale of 0.7, 1, or 1.3. The experiments are conducted on the KTH database under the LOPO setup. (a) Comparison of precision; (b) Comparison of recall.

TABLE I
CONFUSION MATRIX OF SETUP 1. *wlk*: WALKING, *jog*: JOGGING, *run*: RUNNING, *box*: BOXING, *hcp*: HAND CLAPPING, *hwv*: HAND WAVING. EACH ROW REPRESENTS INSTANCES IN AN ACTUAL CLASS, AND EACH COLUMN A PREDICTED CLASS.

	wlk	jog	run	box	hcp	hwv
wlk	**83.3**	13.9	2.8	0	0	0
jog	11.1	**88.9**	0	0	0	0
run	0	27.8	**72.2**	0	0	0
box	0	5.6	0	**83.3**	2.8	8.3
hcp	0	8.3	0	0	**88.9**	2.8
hwv	2.8	2.8	0	0	8.3	**86.1**

TABLE II
CONFUSION MATRIX OF SETUP 2. *wlk*: WALKING, *jog*: JOGGING, *run*: RUNNING, *box*: BOXING, *hcp*: HAND CLAPPING, *hwv*: HAND WAVING. EACH ROW REPRESENTS INSTANCES IN AN ACTUAL CLASS, AND EACH COLUMN A PREDICTED CLASS.

	wlk	jog	run	box	hcp	hwv
wlk	**94.0**	5.0	1.0	0	0	0
jog	1.0	**89.0**	10.0	0	0	0
run	0	14.0	**86.0**	0	0	0
box	0	0	0	**93.9**	4.0	2.0
hcp	1.0	0	0	1.0	**94.9**	3.0
hwv	0	0	0	0	4.0	**96.0**

TABLE III
COMPARISON OF RECOGNITION RATE UNDER THREE EXPERIMENTAL SETUPS. NOTE THAT THE RATES IN PARENTHESIS (KIM *et al.* [18] AND WONG *et al.* [36]) ARE OBTAINED WITH VIDEO SEQUENCES SEGMENTED AND ALIGNED BY HAND. WHILE IN OUR WORK, THE QUERY CLIP IS CROPPED FULLY AUTOMATICALLY AND THE REFERENCE VIDEOS ARE UNSEGMENTED.

Methods	Setup 1	Setup 2	Setup 3
Our method	**83.79**	**92.31**	**92.09**
Jhuang *et al.* 2007 [16]	-	-	91.7
Kim *et al.* 2007 [18]	-	(95.33)	-
Wong *et al.* 2007 [36]	-	71.83 (91.6)	-
Nienles *et al.* 2006 [24]	-	81.50	-
Dollár *et al.* 2005 [9]	-	81.17	-
Ke *et al.* 2005 [17]	62.96	-	-
Schuldt *et al.* 2004 [30]	71.72	-	-

TABLE IV
DETAILED COMPARISON OF RECOGNITION RATE UNDER SETUP 3. *Avg* IS THE AVERAGE PERFORMANCE ACROSS 4 SCENARIOS.

Methods	s1	s2	s3	s4	Avg
Our method	95.56	87.41	90.66	94.72	**92.09**
Jhuang *et al.* 2007 [16]	96.0	86.1	89.8	94.8	91.7

performance is based on the average of five random splits. This is done separately for each scene ($s1$, $s2$, $s3$, or $s4$).

We classify each testing sequence as one of the 6 action types by 3-NN (nearest neighbor). Given a testing sequence, a typical 3D window with 1 second length is cropped out as the query video (as we did in Section V-B). Then the query video is searched against each training sequence, and the maximum value of the similarity surface is viewed as the similarity between the testing and training sequence. The final action type of the testing sequence is determined by voting among the 3 nearest neighbors.

Tables I and II show the confusion matrices for Setup 1 and 2. The results using HSTM are on par with the current state-of-the-art results. Table III lists the comparison of recognition rates by different methods. Our algorithm outperforms the previous methods [9], [16], [17], [24], [30], [36]. Kim *et al.* [18] report a result slightly higher than ours under experimental Setup 2. But they **manually** align the actions in space-time, while our approach is fully automatic. Table IV gives a detailed comparison of recognition rates for each scene ($s1$, $s2$, $s3$, or $s4$) under Setup 3. Our approach achieves an average performance slightly better than that in [16].

D. Run Time of the System

To evaluate the running speed of our approach, we measure the run time in the experiment of querying person smiling in Section V-A. The reference video contains 800 frames (or 26.67 seconds when the frame rate is 30fps) of 200×150 pixels. The query video of a person smiling consists of 32 frames with a resolution 82×106 pixels. The MATLAB program runs on an Intel 3.2Ghz machine. Table V lists the run time for building the HSTM model (only S1 layer, C1 layer, and 3D integral histograms), searching on a single layer (C2, S2, and C1), and coarse-to-fine searching from C2, S2, to S2 layer. We give the absolute time in second as well as the time relative to the video length. The

coarse-to-fine searching process takes about $1/9$ of video time to scan the entire video.

According to Table V, most of the run time is taken by construction of the S1/C1 layers and integral histograms (mainly taken by the C1 layer). Fortunately, these parts of the model need to be pre-computed only once and can be stored for future search. The S2/C2 features are computed online. The run time for this computation is counted as part of searching cost in Table V. Coarse-to-fine searching across multiple layers (from C2, S2 to C1) takes about $1/9$ of video time, nearly as fast as to searching on the C2 layer only. However, without the coarse-to-fine search scheme, the entire C1 layer should be scanned to obtain the same result. This will take nearly 9 times of video time, *i.e.*, resulting in a search speed of 80 times slower. Fig. 11 gives an illustration.

Table V shows that the coarse-to-fine search scheme improves search efficiency. But this is not the only advantage. This scheme also enables the separation of human actions with a wide range of varying motion magnitudes by verifying the candidate locations through different combinations of layers. Also note that both HSTM model and coarse-to-fine searching could be easily implemented using multi-threading and parallel-processing techniques, because most of the computation is convolution and window-scanning. It is expected that parallel-processing will significantly improve the speed.

As to the space requirement, the required memory in the searching stage is very small. We only need to buffer in the

Fig. 11. Run time of searching with/without coarse-to-fine scheme. Coarse-to-fine strategy increases the search speed by a factor of about 80 times.
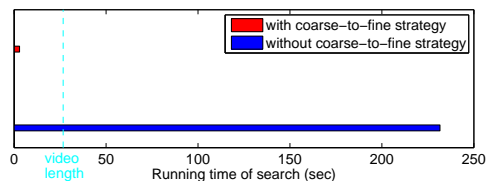
TABLE V

Run time for querying smiling expressions, including the time for building the S1/C1 layers and integral histograms of the HSTM model (*model*), searching on a single layer C2 (*sC2*), S2 (*sS2*), or C1 (*sC1*), and coarse-to-fine (*sC2F*) searching from C2, S2 to C1 layer. The second row is the time in second, and the third row is the second row divided by the video length (*vlen*=26.67 seconds).

	model	*sC2*	*sS2*	*sC1*	*sC2F*
time (*sec*)	2629	2.50	90.06	231.1	2.96
time (*vlen*)	77.3	0.094	3.38	8.66	0.111

memory the query clip and current segment of the reference video. We do need to store to disk the preprocessed intermediate layer for fast search. Storing the C1 layer is optional, depending on the motion magnitudes of actions as we discussed in Section IV (*e.g.*, it is unnecessary to store the C1 layer of the KTH dataset.). Due to down-sampling, the C1 layer size is $1/32$ of the original size (see Section III). Considering that the C1 layer has 9 orientations and 2 scales and that the original video has 3 color channels, the C1 layer requires about 20% of storage space of the original uncompressed video. The intermediate layer consists of integral histograms of the C1 layer. Hence, it requires the same storage space as the C1 layer. Note that both C1 layer and intermediate layer could be compressed to largely reduce the storage size, without affecting performance too much.

VI. Discussions

Our model makes selectivity and invariance of feature characteristics coexist in the hierarchy. As holistic histograms, the C2 features are very invariant but lack selectivity. The S2 features (HIGO) have high invariance, because the histograms are computed over small cells, while it is also selective because the geometric information is kept by maintaining the order of the cells. The C1 feature is very selective for differentiate subtle motions, and its invariance comes from pooling over regions. This property of our model enables separation of actions ranging from rapid sports movements to subtle facial expressions by stopping coarse-to-fine search on a certain layer. The ballet, smiling, as well as other sequences demonstrate this capability of our model. The capability of tolerating variations is further verified on the KTH dataset. KTH contains large variations in view angle, spatial scale, appearance, human body shape, *etc*. We also add temporal variations by interpolating query clips in time axis. Our model can differentiate most of the video sequences by action type while being resistent to the variations. Note that the coarse-to-fine search is stopped at the S2 layer for the KTH dataset. It implies that the S2 feature (HIGO) has good trade-off between selectivity and invariance in differentiating human body movements.

Both Jhuang's [16] and our model represent motion observations using alternative "S" and "C" layers, borrowing some ideas from biological models [14], [15]. But the basic purposes are quite different. We aim to search actions of interest in large video databases by querying short clips, while Jhuang *et al.* [16] focus on recognizing action types of testing sequences. These two vision tasks differ greatly in that, recognition aims to build classifiers with good generalization ability, while action search requires high efficiency of correlating two videos and capability of

handling completely new action categories. This leads to different hierarchical models (see more detail in Section II). Although our model can also be modified for action recognition and even achieves performance better than Jhuang's [16] work, we want to emphasize the advantages of our model on action search that includes high efficiency and accuracy, and capability of handling varying motion magnitudes (ranging from rapid sports movements to subtle facial motions). Jhuang's model might also be able to work on action search after major modifications. But further experiments are needed to test its performance and capability of handling new action categories.

Although our approach works well for short query clips ($1-2$ seconds), we did not test it on longer queries of actions with changing rhythms. In that case, the dynamic time warping may help to improve the performance of matching. Another limit of our approach is that it may fail when the video data have very large variance in scale and view angle, although it preserves feature selectivity and invariance to some extent. A pyramid of the features might solve this problem but at the expense of computational cost. We leave these for our future work.

VII. Conclusions

We propose a hierarchical space-time model (HSTM) for searching human actions in videos, inspired by the biological model of visual processing in primate cortex [14], [15] and the quantitative models (Neocognitron [11], CNN [21], and HMAX [28]) for object recognition. Simulating the simple and complex cells in the biological model, the HSTM has 5 layers, which, ranging from low layer to high layer, introduces a novel set of features for videos with increasing invariance and decreasing selectivity. This model naturally leads to an efficient coarse-to-fine search strategy that enables the search of human actions ranging from rapid sports movements to small motions as subtle as facial expressions. The computational efficiency of the HSTM model makes it possible to search human actions in large video databases (*e.g.*, the KTH database). The claim that the HSTM model is both selective and robust for searching human actions is validated by our experimental results.

References

[1] http://www.ifp.uiuc.edu/~ hning2/hstm.htm.
[2] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. *ICML*, 2004.
[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.
[4] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *ICCV*, 2005.
[5] O. Chomat and J. L. Crowley. Probabilistic recognition of activity using local appearance. *CVPR*, 02:2104, 1999.
[6] O. Chomat, J. Martin, and J. L. Crowley. A probabilistic sensor for the perception and recognition of activities. In *ECCV (1)*, pages 487–503, 2000.
[7] N. P. Cuntoor and R. Chellappa. Epitomic representation of human activities. *CVPR*, 2007.
[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.
[10] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003.
[11] K. Fukushima. Neocognitron: A self-organizing neural network model for a machanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.
[12] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4:179–192, March 2003.

[13] C. Harris and M. Stephens. A combined corner and edge detector. *In Alvey Vision Conference*, pages 147–152, 1988.

[14] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cats visual cortex. *J. Physiol. (Lond.)*, 160:106–154, 1962.

[15] D. Hubel and T. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol*, 28:229–289, 1965.

[16] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *ICCV*, 2007.

[17] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *International Conference on Computer Vision*, volume 1, pages 166 – 173, October 2005.

[18] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. *CVPR*, 2007.

[19] S. Kullback. *Information Theory and Statistics*. Dover Publications, New York, 1968.

[20] I. Laptev and T. Lindeberg. Space-time interest points. volume 01, page 432, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[23] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *CVPR*, 2006.

[24] J. Niebles, H. Wang, H. Wang, and L. Fei Fei. Unsupervised learning of human action categories using spatial-temporal words. *BMVC*, page III:1249, 2006.

[25] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. *CVPR*, 2007.

[26] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *CVPR*, 2005.

[27] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. *NIPS*, 2004.

[28] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2, 1999.

[29] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[30] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. *ICPR*, 2004.

[31] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *PAMI*, 29 (3):411–426, 2007.

[32] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005.

[33] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR*, 2005.

[34] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.

[35] N. Vaswani, A. RoyChowdhury, and R. Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. *CVPR*, 2003.

[36] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. *CVPR*, 2007.

[37] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. *ICCV*, 2005.

[38] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, pages 123–130, 2001.

Huazhong Ning received the BSc degree in computer science from University of Science and Technology of China, Hefei, China in 2000, and received the MSc degree in pattern recognition and intelligence systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China in 2003, and is currently working toward the PHD degree in electrical engineering at University of Illinois at Urbana-Champaign. He has worked as a 3G Software Engineer in Alcatel Shanghai Bell, China and as summer interns in NEC American Labs, USA. His current research interests include video/image processing, machine learning, clustering, audio analysis, data mining, *etc*.
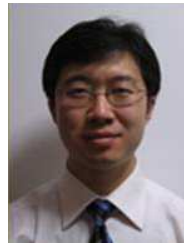
Tony X. Han received the B.S. degree with honors in Electrical Engineering Department and Special Gifted Class from Jiaotong University, Beijing, P. R. China in 1998, M.S. degree in Electrical and Computer Engineering from the University of Rhode Island, RI, in 2002, and Ph.D degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, IL, in 2007. He then joined the Department of Electrical and Computer Engineering at the University of Missouri, Columbia, MO, in August 2007. Currently, he is an assistant professor of the Department of Electrical and Computer Engineering. He received CSE Fellowship form University of Illinois in 2005.

Dirk B. Walther is a Beckman Postdoctoral Fellow at the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. After finishing his in Computation and Neural Systems at the California Institute of Technology he worked as a postdoctoral fellow at the Centre for Vision Research at York University, Toronto, Canada, before joining the Beckman Institute. His research interests include perception natural scenes in biological and engineering systems, including interactions of visual attention and object recognition.
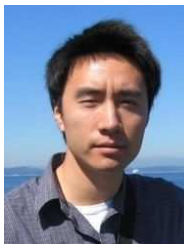
Ming Liu received both BE and ME degrees in electrical engineering from University of Science and Technology of China, Hefei, China in 1999 and 2002 respectively, and Ph.D in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, IL, in 2007. He then joined the Video Search group in Microsoft Corporation. His research interests include speaker/speech recognition, image/video processing, face recognition, and audio/visual fusion via machine learning.

Thomas S. Huang (S'61-M'63-SM'76-F'79-LF-01) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, R.O.C., and the M.S. and D.Sc. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the Faculty of the Department of Electrical Engineering at MIT from 1963 to 1973 and the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University, West Lafayette, IN, from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, Urbana, where he is now the William L. Everitt Distinguished Professor of Electrical and Computer Engineering, a Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology and Co-Chair of the Institute's major research theme Human Computer Intelligent Interaction. He has published 20 books, and over 500 papers in network theory, digital filtering, image processing, and computer vision. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals.