Visual Comput (2008) 24: 595–603
DOI 10.1007/s00371-008-0240-1

ORIGINAL ARTICLE

Yadong Mu
Shuicheng Yan
Thomas Huang
Bingfeng Zhou

Contextual motion field-based distance for video analysis*

Published online: 29 May 2008

Y. Mu (✉)· B. Zhou (✉)
Institute of Computer Science and Technology, Peking University, Beijing 100871, P.R. China
muyadong@gmail.com,
ccbfzhou@pku.edu.cn

S. Yan
ECE Department, National University of Singapore, 117576 Singapore, Singapore

T. Huang
ECE Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Abstract In this work, we propose a general method for computing distance between video frames or sequences. Unlike conventional appearance-based methods, we first extract motion fields from original videos. To avoid the huge memory requirement demanded by the previous approaches, we utilize the "bag of motion vectors" model, and select Gaussian mixture model as compact representation. Thus, estimating distance between two frames is equivalent to calculating the distance between their corresponding Gaussian mixture models, which is solved via earth mover distance (EMD) in this paper. On the basis of the inter-frame distance, we further develop the distance measures for both full video sequences.
Our main contribution is four-fold. Firstly, we operate on a tangent vector field of spatio-temporal 2D surface manifold generated by video motions, rather than the intensity gradient space. Here we argue that the former space is more fundamental. Secondly, the correlations between frames are explicitly exploited using a generative model named dynamic conditional random fields (DCRF). Under this framework, motion fields are estimated by Markov volumetric regression, which is more robust and may avoid the rank deficiency problem. Thirdly, our definition for video distance is in accord with human intuition and makes a better tradeoff between frame dissimilarity and chronological ordering. Lastly, our definition for frame distance allows for partial distance.

Keywords Video analysis · Motion field · Activity classification

1 Introduction

In recent years, analysis of video information has attracted growing attention from the computer vision community. This is mainly because videos provide much more information than separate images and therefore we have the possibility to make many vision tasks practical. Related applications include event detection [1, 9, 14], event clustering [22], action classification [1], and others [10].

One of the crucial issues in the above topics is how to estimate the distance or similarity between two frames or videos, which should reflect similitude of video contents, especially regarding motions or behaviors. The video appearance provides few cues due to the fact that the same behavior could have different spatial properties in videos (for example, two people wearing different clothes perform the same action). On the contrary, in-

*This work was supported by China NSF Grant No. 60573149, Beijing NSF Grant No. 4072013.

tensity gradients and optical flow are more robust and reasonable. Many motion-based similarity measures have been proposed in the recent literature. While in some approaches parametric models and specific types of events are needed [4], there are others designed for general purpose [9, 22].

In this paper, we aim at providing a general approach to the above problem based on video motion patterns. The works most related to ours are found in [4, 14, 22]. According to our observation, there are two distinct ways to utilize video motions: space-time gradient-based methods [22] and optical-flow-based methods [4, 9]. Although the former proved effective in event-based analysis of video in Zelnik and Irani's work [22], we argue that optical flow is more essential. Note that optical flow roughly indicates an object's motion direction. It could form a tangent vector field of the 2D spatio-temporal surface manifold generated by the motion event contained in video. It is known that in the ideal case optical flow should be normal to possible intensity gradients. For a space-time point there could be multiple feasible gradients even its optical flow is uniquely decided. Thus, optical flow is more fundamental to describe video motions.

There are many ways to estimate the similarity between two objects (signal, images, shapes, etc.), among which Boiman and Irani's criterion [2] is one of the most interesting. The idea behind this work is reasonable: if an object $S1$ could be relatively easily composed by different parts of another object $S2$, then there is a high possibility that they share a large similarity score, which is called *similarity by composition* [2] and several examples for illustration are found in the work. In this paper we extend this idea to the field of video analysis. Our main contribution lies in introducing the transportation distance [12] to provide reasonable solutions to the *similarity by composition* idea. Specifically speaking, information in a single video frame is explicitly clustered into several distinct parts, and the distance measure between two frames is proportional to labors spent on transforming one into the other, which could be seen as an equivalent expression of the above-mentioned composition operation.

2 Overview

It is the common case to model video as separate images and process videos in a frame-by-frame manner, which has been proven effective in many applications as shown in [19]. However, taking into account the inter-frame correlations would go one step further towards fully exploiting the information contained in videos. In this paper, we model video sequences using dynamic conditional random field (DCRF, see [18] for formal definition), which is a graphical model (see Fig. 1) satisfying the Markov property typically with a one-dimensional chain topology. Each node in it is conditioned on the whole data and also

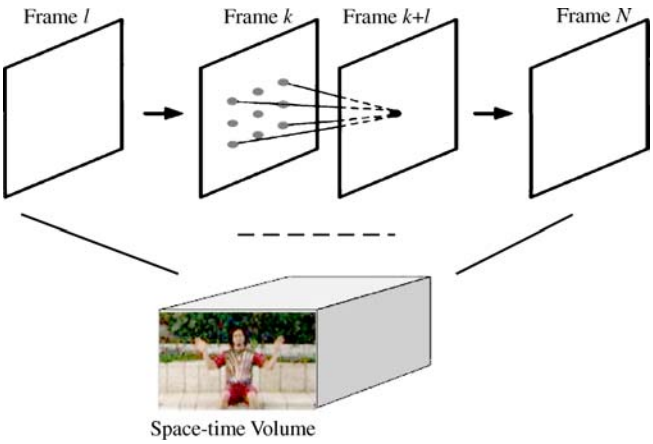


Fig. 1. Graphical model for DCRF

a Markov random field. The DCRF formulation relaxes our burden to model the complicated dependencies in source data, and seriously regards influence between adjacent graph nodes along all dimensions. Under the DCRF formulation, the motion vector for each frame pixel only depends on its small local neighbor patch (both in the spatial and temporal sense).

It is another crucial issue to decide how to model motion information in a single frame. For simplicity and efficiency, we ignore the spatial configuration information within one frame and model it by "bag of motion vectors". We represent motion fields with the Gaussian mixture model (GMM), which is proved to be a fairly efficient and compact representing form to approximate the real probability distribution. The initial parameters for GMM are computed through k-means, and are subsequently refined by iterative expectation-maximization (EM) [3].

Estimating the dissimilarity between two frames is reasonable and straightforward by comparing their corresponding GMMs. As argued in [7], the distance for GMMs could be expressed in the Kullback–Leibler (KL) divergence form. However, as KL-divergence between two GMMs could not be analytically computed, in implementation we approximate it by the earthmover distance similar to [6]. Keep in mind that this is only the distance between frame pairs. For two integrated video clips, we have to calculate their similarity basing on the frame-to-frame distances. From our observations, it could be considered to be a many-to-one frame matching problem. A good matching between two input video sequences should respect both the dissimilarity between matching frame pairs and inner chronological orders. Efros's patch correlation-based method [4] results in a rigid temporal consistency. A similar issue arises in [22], where a sliding window shifts across the entire sequence. As an alternative, we define an energy function that could tolerate more temporal inconsistency, and search for its global optimal using belief propagation (BP) [5, 15].

3 Motion field

3.1 Probabilistic modeling for optical flow fields

It is non-trivial to model motion fields using a probabilistic language, since more advanced statistical methods could be potentially introduced for refinement. We treat a video sequence as a space-time volume following chronological order. For each image pixel, a 3-dimensional motion vector (u, v, w) is associated to it. If videos are captured at a constant time interval, the velocity components projected on the time coordinate are always equal to 1. Hence our aim is reduced to estimate the 2-tuple (u, v), which could be roughly understood as an object's respective offset on the x and y coordinates relative to the previous frame.

With the assumption that a pixel is only correlated with its space-time neighbors, we model the video sequences by dynamic conditional random fields (DCRF) [18], whose graphical model is illustrated in Fig. 1. Let M and Z denote the motion fields and space-time volume with RGB channels, respectively. m_p is the motion vector for pixel p. According to the Hammersley–Clifford theorem [11], the negative log form of posterior could be represented as below:

$$\begin{aligned} E &= -\log P(M|Z) \\ &= -\sum_{p} \log \phi_1(m_p|Z) - \sum_{y\in N_p} \log \phi_2(m_p, m_y|Z) \\ &\quad - \sum_{q_i\in N'_p} \log \phi_3(m_p, m_{q_1}, \ldots, m_{q_k}|Z) + \text{const.}, \end{aligned} \tag{1}$$

where N'_p and N_p both are neighborhood systems, yet differ in that the former is a multi-variable function which describes the relation to neighbors in previous or subsequent frames, while the latter is binary and located in the current frame. Note that motion vectors here all have continuous real values, which complicates the computation. Current estimating methods for optical flow are mostly developed based on the first two terms in Eq. 1. However, to utilize contextual information is beneficial and brings about more accurate and consistent estimation in many cases. In the following subsection, we developed *Markov volumetric regression* to explicitly exploit this kind of information.

3.2 Markov volumetric regression

For videos captured with a high noise level, the estimated motion field is corrupted by small perturbations and noises. Previous optical flow algorithms, such as classical Lucas–Kanade (LK) [16], operate in a frame-by-frame style and usually adopt a special denoising strategy for noisy image sequences. However, contextual information contained in chronologically adjacent frames could provide other valuable clues to correct erroneous estimation.

To reduce the noise level in low-speed motions, more consistency within a temporal neighborhood is favorable. However, it is difficult to describe a well-defined criterion for the consistency. We adopt an idea similar to that of [17] and [20], where the authors argue that a good estimation for a current pixel should be approximately inferred from its neighbors. This idea is called *local consistency* in some literature. For clarity, first we will briefly review the optical flow equation, and then illuminate how the idea of local consistency benefits.

For each space-time point q positioned at (x, y, t) and inside the 3D patch around pixel p, the partial derivative of video intensity I on time t satisfies the optical flow equation [16]:

$$\frac{\mathrm{d}I_q}{\mathrm{d}t} = \frac{\partial I_q}{\partial x}u + \frac{\partial I_q}{\partial y}v + \frac{\partial I_q}{\partial t}w = \nabla I_q \cdot m_q = 0, \tag{2}$$

where $\cdot$ denotes the inner product between intensity gradient and motion vector. The inner product equal to 0 reflects the orthogonality of two vectors.

Now we can begin to detail the proposed *Markov volumetric regression* procedure. Our method operates on 3-D space-time patches around current position. In implementation we typically select a $5 \times 5 \times 3$ patch, which amounts to the fact that current estimation is affected by both frames before and after it, rather than either of the two alone. This brings a stronger smoothing effect. The spatial and temporal extents of the small patch could be adjusted according to video resolution and motion velocity. If we pile all the constraints derived from Eq. 2 in a matrix form, we can obtain the following representation:

$$A\begin{pmatrix} u \\ v \end{pmatrix} = -b, \tag{3}$$

where

$$A = \begin{pmatrix} \frac{\partial I_{q_1}}{\partial x} & \frac{\partial I_{q_1}}{\partial y} \\ \vdots & \vdots \\ \frac{\partial I_{q_k}}{\partial x} & \frac{\partial I_{q_k}}{\partial y} \end{pmatrix} \quad b = \begin{pmatrix} \frac{\partial I_{q_1}}{\partial t} \\ \vdots \\ \frac{\partial I_{q_k}}{\partial t} \end{pmatrix}.$$

Following the intuition that pixels in the previous frame with a similar color might have a big chance to share the same motion patterns, we can put extra local consistency constraints into Eq. 3, which implies the Markov property of the motion fields. That is:

$$A^* = \begin{pmatrix} A \\ \lambda & 0 \\ 0 & \lambda \end{pmatrix} \quad b^* = \begin{pmatrix} b \\ \lambda \sum_q w_q u_q \\ \lambda \sum_q w_q u_q \end{pmatrix},$$

where $w_q \propto \exp(-\|z_p - z_q\|^2/\sigma_c^2)$, and z denotes the color vector of a pixel. The sum of all w_q should be normalized to be equal to one. The motivation for in-

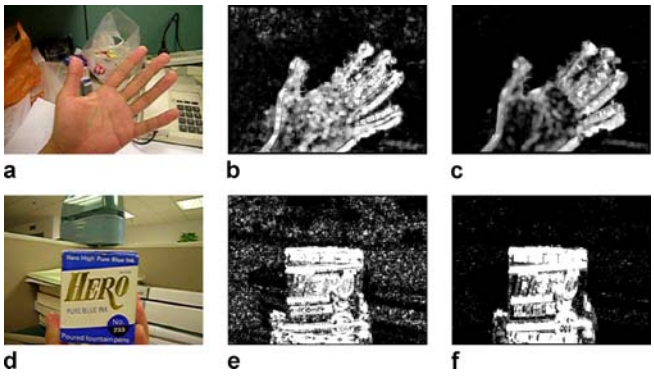


Fig. 2a–f. Comparison of Markov volumetric regression (*right column*) and Lucas–Kanade algorithm (*middle column*). Brightness is proportional to motion velocity. Note that the proposed regression method suppresses background noises and tends to smoother motion estimations compared to LK

troducing extra terms in Eq. 3 could be summarized as: adjust its own estimation according to the received contextual information and propagate labels or values across its neighborhood. λ reflect our belief for the consistency with the previous estimation, which could be user-specified or automatically adjusted. Also note that in the new equation, the two columns of matrix A^* are always linear independent, thus avoiding rank deficiency [8]. A comparison with the LK algorithm could be found in Fig. 2. Note that both noises are suppressed and motion boundaries are thickened due to smoothing.

4 Distance measure for frames

4.1 Motion representation for a single frame

Directly storing motion fields obtained in previous steps needs huge memory, thus a compact and effective representation should be seriously chosen. The Gaussian mixture model (GMM) is one of the most popular models suitable for such a task. We firstly cluster the input data according to motion similarity and spatial adjacency (in current context, they form a 4-tuple vector, two for motion and two for position), and then represent only using the motion components of its cluster center, since position and size of an object are typically correlated with spatial resolution and usually most effective only when their parametric models are known and well-defined as guidance.

The clustering process starts with k-means to get initial segments. Then iterative EM [3] is taken to optimize Gaussian parameters. Finally, we get several clusters, with their priors, means, and variances known, i.e. $\{w_i, (u_i, v_i), \sigma_i, i = 1 \ldots K\}$, where K is the predefined cluster numbers. Two illustrative examples are found in Fig. 3.

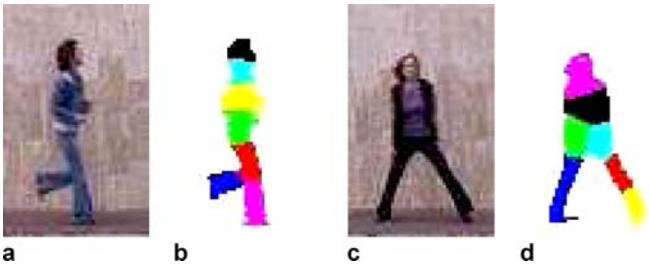


Fig. 3a–d. GMM representation for motion fields. Motion vectors are clustered according to motion similarity and spatial adjacency. Pixels that belong to the same cluster are in one color

4.2 Earthmover-based frame distance

There are some existing approaches proposed to measure distance between two GMMs [6, 7], among which an interesting one is defined in a KL-divergence sense [7]. However, KL-divergence is not appropriate for our case for its computation complexity. Instead we adopt the earthmover distance (EMD) [12, 13] as the distance measure. The choice of EMD is not only for its efficiency, but also for other attractive characteristics. In the following, first we will briefly introduce EMD's principles, and then list its intuitive meanings for our application and advantages over other alternatives.

Earthmover distance could be best understood via the following analogy: for two sets of weighted features, one could be viewed as "earth", while the other as "holes" in a corresponding place. Formally speaking, let $S = \{(s_1, w_{s_1}), \ldots, (s_m, w_{s_m})\}$ and $T = \{(t_1, w_{t_1}), \ldots, (t_n, w_{t_n})\}$ denote two distributions consisting of distinct features and corresponding weights or priors. A ground distance needs to be defined to evaluate a unit of work for transporting a unit of earth to the holes. Let $\mathrm{dist}(s, t)$ denote the ground distance between two features s and t. The EMD between S and T is then given as below:

$$\mathrm{EMD}(S, T) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \mathrm{dist}(s_i, t_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}, \tag{4}$$

where $f_{ij} \geq 0$, represents the optimal admissible flow from s_i and t_j. The calculation of EMD in Eq. 4 is equivalent to seeking an optimal f_{ij} subject to the following constraints:

$$\sum_{j=1}^{n} f_{ij} \leq w_{si}, \quad \sum_{i=1}^{m} f_{ij} \leq w_{tj}$$
$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min\Big(\sum_{i=1}^{m} w_{si}, \sum_{j=1}^{n} w_{tj} \Big). \tag{5}$$

For the current task, we ignore the variances in GMM, and treat each mean as a feature, prior as its weight. The

ground distance in Eq. 4 could have various forms. The most frequently used distances are Euclidean and cosine. While the former depends on the magnitude of difference between two vectors, the cosine distance is determined by motion direction only and the effect of motion magnitude is removed carefully. This treatment is based on the following observation: the same "walking" action taken by an old man or a teenager may be classified to different categories since they move fast or slow, which is certainly against the true situation. In the following we will use different distances according to the requirements of different applications. Given two means $\bar{m}_1$ and $\bar{m}_2$, their cosine distance could be calculated as:

$$\mathrm{dist}(\bar{m}_1, \bar{m}_2) = 1 - \frac{\bar{m}_1 \cdot \bar{m}_2}{\|\bar{m}_1\| \|\bar{m}_1\|}. \tag{6}$$

The most attractive feature of EMD is its accordance with human intuition. For example, in Fig. 4 we segment three video frames into small parts, each of which undertakes a uniform motion. It is obvious to find out that the first two have similar motion patterns compared to each other, while this is not true for the third frame. In fact, it is effortless for a human to search for matching pairs between the two sets of moving parts. Intuitively, a distance definition basing on the extent of how they match each other seems fairly reasonable. Fortunately, EMD is qualified for this task. EMD distance is a relatively accurate description of the effort required to compose one object from distinct parts of another, thus is superior to many other distance definitions in this aspect.

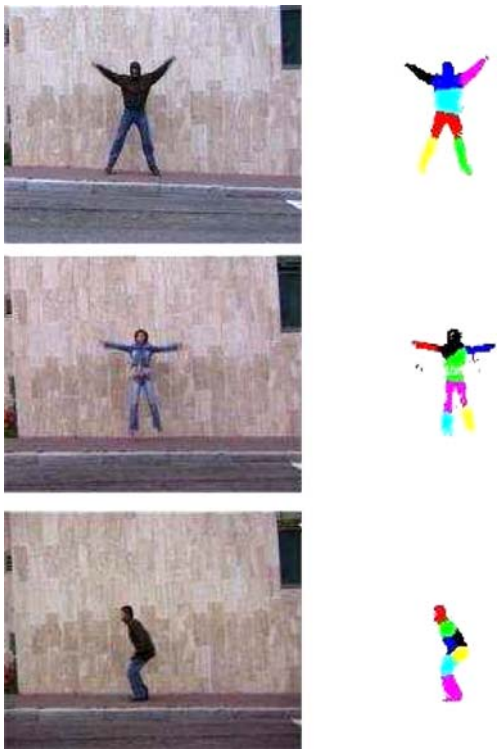


Fig. 4. Intuitive explanation for EMD. The *left column* shows frames selected from three videos, and the *right column* shows the segmentations according to motion similarity and spatial adjacency

Our earthmover-based definition has other merits. Here we describe its usefulness for the computation of partial distance. In the case of multiple motions, blindly computing between the entire images sometimes fails to disclose the truth. For example, the "waving" and "rotating" videos in Fig. 5 could be seen as parts of the "waving and rotating" video (the leftmost video), thus are supposed to have large similarity. However, directly applying Eq. 4 does not reflect their true similar extents. Substituting pixel counts for the Gaussian weights in GMMs could utilize the power of EMD to calculate partial distance, namely skipping the normalization step in the EM algorithm. Consequently, certain components and their counterparts share similar weights, thus lower distances are obtained. A simple illustration of the above idea is presented in Fig. 6. For the purpose of evaluation, given two video sequences $V1 = \{S_i\}$,

Fig. 5a–c. Videos for partial distance illustration. **a** Video 1 – "waving and rotating". **b** Video 2 – "waving". **c** Video 3 – "rotating"

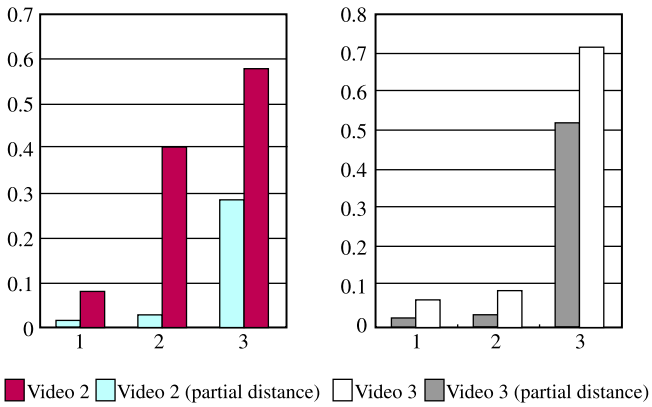


Fig. 6. Experimental results for partial distance. We calculate both the distances of Video 2 and Video 3 to Video 1 in Fig. 5 with or with out partial consideration. The *left column* is for the "waving" video, and the *right column* for the "rotating" video. The three *bins* in the horizonal coordinate denote the statistics $S1$, $S2$ and $S3$ (see Eqs. 7–9), respectively. Note that $S2$ for Video 2 is drastically reduced due to partial distance, which could be intuitively understood via the idea of similarity by composition

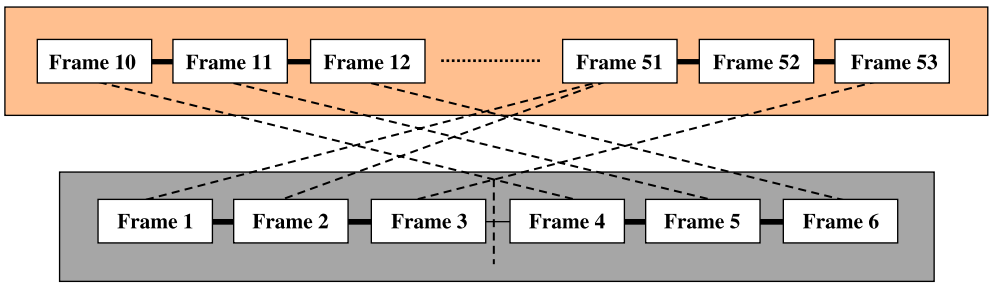


Fig. 7. Example for frame matching with belief propagation. Frame neighbors are connected with a *solid line*, whose width is proportional to the dissimilarity extent. The vertical *dashed line* between frame 3 and 4 in the source video (the *lower* one) shows the fact that they are very dissimilar from each other thus could automatically break the chronological constraints here

$i = 1, \ldots, L$ and $V2 = \{T_j\}$, $j = 1, \ldots, K$, we define three statistics for comparing the partial distance with the original one:

$$S1 = \frac{1}{L} \sum_{i=1}^{L} \min_j \mathrm{EMD}(S_i, T_j) \tag{7}$$

$$S2 = \frac{1}{K} \sum_{j=1}^{K} \min_i \mathrm{EMD}(S_i, T_j) \tag{8}$$

$$S3 = \frac{1}{LK} \sum_{i=1}^{L} \sum_{j=1}^{K} \mathrm{EMD}(S_i, T_j). \tag{9}$$

5 Distance measure for videos

After defining distance between two frames, it is now possible to estimate how similar two video sequences are. As argued before, a good distance definition should make a tradeoff between dissimilarities for matching frame pairs and consistency with the chronological orders. Here we define an energy function satisfying the above requirements:

$$E = \lambda_v \sum_{p,q} w_{pq} \exp\left(-\|(f_p - f_q) - (p - q)\|^2 / \sigma_v^2\right) + \sum_{i=1}^{L} \mathrm{EMD}(S_i, T_{f_i}), \tag{10}$$

where p and q are indices for adjacent frames, f_i denotes the matching frame's index in target video T for frame i in source video S. U is the upper bound for EMD. $w_{pq} \propto (U - \mathrm{EMD}(S_p, S_q))$ and should be normalized and scaled. The first term reflects the temporal-consistency constraints, while the second term is the sum of all matching frame pairs in source and target videos. λ_v is a positive constant introduced to balance two energy terms in Eq. 10.

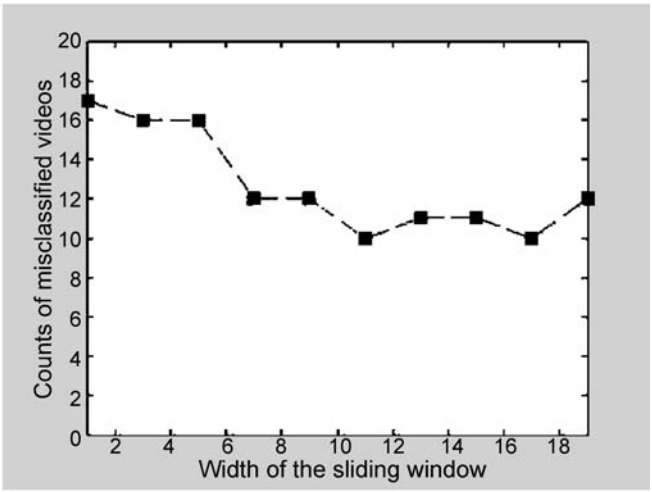
Finding the optimal matching on the whole video sequences has high complexity both in computing time and memory. Typically we firstly divide the source video into several segments with similar motions and a relatively shorter length than the target video (possible segmenting strategies include greedy algorithm, clustering techniques discussed in Sect. 6, or other advanced schemes), and search for optimals using belief propagation (BP) [5, 15].

Conventional methods such as dynamic time warping (DTW) and sliding window (see [4]) have their limitations for the optimization of Eq. 10, since they handle chronological ordering in a too rigid way. On the contrary, BP could handle situations like in Fig. 7 more flexibly, where the matching frames of the last three frames lie before the first three, while within each group chronological orders are kept. Moreover, it could be proved that sliding window is a special form of BP. Namely, when λ_v in Eq. 10 tends to infinity, BP is equivalent to sliding window.

Fig. 8. Weizmann video database for action classification experiment

	Bend	Jack	Jump1	Jump2	Run	Side	Skip	Walk	Wave1	Wave2
Bend	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Jack	.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00
Jump1	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00
Jump2	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00
Run	.00	.00	0.20	.00	.70	.00	.10	.00	.00	.00
Side	.00	.00	.00	.00	.00	0.78	.00	0.22	.00	.00
Skip	.00	.00	.50	.00	.00	.00	.50	.00	.00	.00
Walk	.00	.00	.00	.00	.00	.00	.00	1.00	.00	.00
Wave1	.00	.00	.00	.00	.00	.00	.00	.00	1.00	.00
Wave2	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00

a



b

Fig. 9. a Confusion matrix for Weizmann video database (Performance average is 89.25%). **b** Curve for the misclassified videos

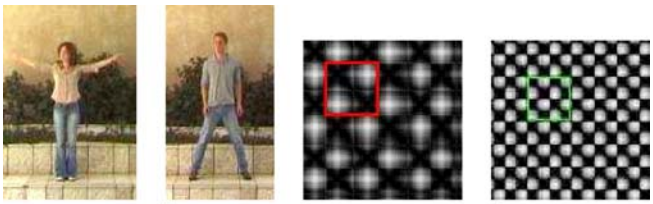
6 Experiments

Potential applications of our proposed distance measure include human activity classification, unsupervised and semi-supervised video clustering, event detection etc.

1. Activity classification. Given a database of example human activities, we can judge the category for a new frame or video, basing on the distance previously defined. Examples are shown in Figs. 8 and 9. There are in total ten kinds of distinct human motions and the whole Weizmann database[1] consists of 93 video clips. For each video we perform a leave-one-out procedure, i.e. we treat the other 92 videos as the training set. A confusion matrix is given in the left of Fig. 9 and the relations between error classification and sliding window size (for convenience of comparison we set λ in Eq. 10 to infinity) is curved in the right. The algorithm misclassified 10 of 93 videos, which is relatively higher than the original results in [1]. This is mainly because we dropped spatial configurations for efficiency and no preprocessing steps such as background subtraction in [1], also there are no user interactions. However, this makes our algorithm more general, not limited to static camera and low-level noises as in [1]. Information like spatial moments is supposed to be beneficial for higher accuracy.

2. Unsupervised video clustering. We can also perform clustering tasks basing on the *distance matrix* $D = \{d_{ij}\}$, where d_{ij} is the EMD distance between frame i and j. Examples are shown in Figs. 10 and 11. For periodical behaviors, we finally get chessboard-like distance matrices and it is convenient to locate one cycle for specific motion patterns (see Fig. 10). For videos with isolated actions, the distance matrices are much more complicated. In Fig. 11, we calculate the distance matrix of the "Stefan" video with itself, and with "walking" and "jumping" videos. We select four representative frames from "Stefan" for a better understanding of the distance matrix. As seen in Fig. 11, the motion-based clustering results for "Stefan" are obvious. Current clustering methods, such as normalized cuts [21], or even simple k-means, are qualified for the segmenting task.

Moreover, note that "Stefan" is captured with a fast-moving camera. Although in its phase 1 and 3 the tennis player moves to the right, the dominant motions for current frames are inversely left-moving, which seems a bit strange at first glance. In fact, the camera's movement could account for this. To keep the player in the center of frames, the camera keeps up with the player, thus the au-



a b c d

Fig. 10a–d. Unsupervised clustering analysis for "waving" and "jumping". **c**, **d** are distance matrices for the source videos respectively, and matrix element's brightness is proportional to its distance value. Both of the two videos contain periodical actions. We calibrate one cycle (emphasized in *green* and *red boxes*, respectively). Further analysis for the "jumping" video in **b** could be found in Fig. 12

[1] http://www.wisdom.weizmann.ac.il/~vision

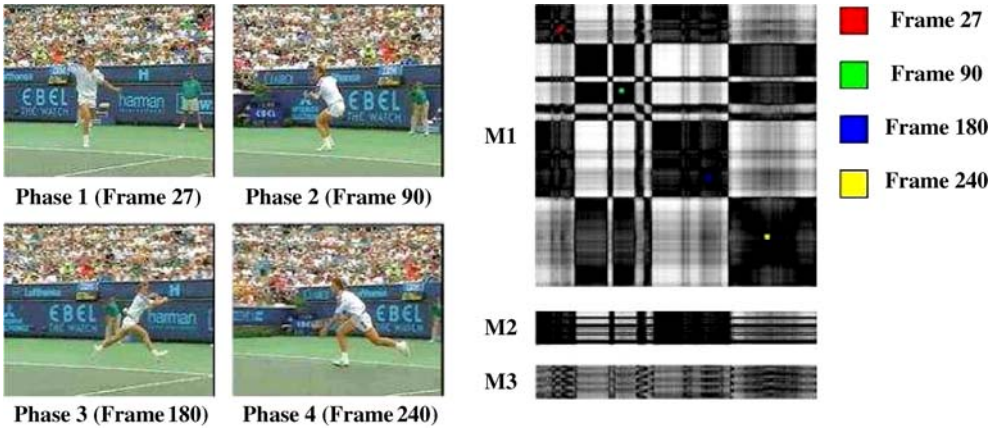


Fig. 11. Distance matrices for "Stefan" standard testing video sequence (*left*). M1: distance matrix for "Stefan" itself. M2: distance matrix for a "walking" video. M3: distance matrix for a "jumping" video

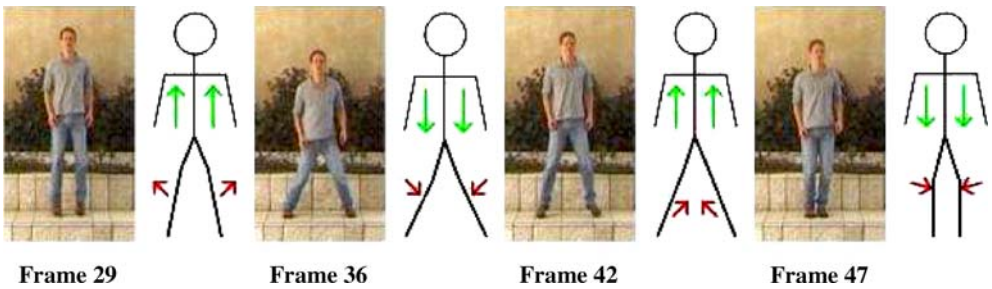


Fig. 12. Motion phase analysis for the "jumping" video. See text for explanation

dience and tennis court seem to undertake a left-moving motion. Thus, phase 1 and 3, rather than the other two phases, are more similar to the "walking" video, which contains a walking action from right to left. The distance matrix for the "jumping" sequence is dominated by bright intensity, which is in accord with the absence of up-down movement in the "Stefan" video.

3. Motion phase analysis. Another interesting application of our method is analyzing motion phases in a video. We take the "jumping" video in Fig. 10 for illustration. The jumping action could be seen as a four-phase procedure rather than two, which is at contrast to many people's intuition. Moving directions are indicated using arrows. The experimental results are illustrated in Fig. 12.

7 Conclusion and future work

In this paper we propose a new kind of motion descriptor for video frames, and rely on it to measure distance between video frames and sequences. Various experiments are presented to exhibit its effectiveness. However, how to refine the motion estimation quality and exploit the spatial configuration in the motion field is less discussed and left for further exploration.

References

1. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, pp. 1395–1402. IEEE Computer Society, Washington, DC (2005)
2. Boiman, O., Irani, M.: Similarity by composition. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience Publication (2000)
4. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV, pp. 726–733. IEEE Computer Society, Washington, DC (2003)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. Int. J. Comput. Vision **70**(1), 41–54 (2006)
6. Greenspan, H., Dvir, G., Rubner, Y.: Context-dependent segmentation and matching in image databases. Comput. Vis. Image Underst. **93**(1), 86–109 (2004)
7. Greenspan, H., Goldberger, J., Ridel, L.: A continuous probabilistic framework for image matching. Comput. Vis. Image Underst. **84**(3), 384–406 (2001)

8. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer (2001)
9. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: International Conference on Computer Vision, vol. 1, pp. 166–173. IEEE Computer Society, Washington, DC (2005)
10. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, pp. 432–439. IEEE Computer Society, Washington, DC (2003)
11. Li, S.Z.: Markov Random Field Modeling in Image Analysis (Computer Science Workbench). Springer (2001)
12. Rubner, Y., Guibas, L.J., Tomasi, C.: The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In: APRA Image Understanding Workshop, pp. 661–668 (1997)
13. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
14. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: CVPR (1), pp. 405–412. IEEE Computer Society, Washington, DC (2005)
15. Sun, J., Yuan, L., Jia, J., Shum, H.Y.: Image completion with structure propagation. ACM Trans. Graph. **24**(3), 861–868 (2005)
16. Trucco, E., Verri, A.: Introductory Techniques for 3-D Computer Vision. Prentice Hall PTR, Upper Saddle River, NJ (1998)
17. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. In: ICML, pp. 985–992. ACM, New York, NY (2006)
18. Wang, Y., Ji, Q.: A dynamic conditional random field model for object segmentation in image sequences. In: CVPR (1), pp. 264–270. IEEE Computer Society, Washington, DC (2005)
19. Winnemöller, H., Olsen, S.C., Gooch, B.: Real-time video abstraction. ACM Trans. Graph. **25**(3), 1221–1226 (2006)
20. Wu, M., Schoelkopf, B.: A local learning approach for clustering. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007)
21. Yu, S.X., Shi, J.: Multiclass spectral clustering. In: ICCV, pp. 313–319, IEEE Computer Society, Washington, DC (2003)
22. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: CVPR (2), pp. 123–130. IEEE Computer Society, Los Alamitos, CA (2001)




YADONG MU received the B.S. degrees from both the Computer Science Department and Philosophy Department, Peking University, Beijing, China, in 2004 and 2005 respectively. Currently, he is a Ph.D. student in the Institute of Computer Science and Technology, Peking University. His research interests include computer graphics, computational photography, computer vision and machine learning.

SHUICHENG YAN received the B.S. and Ph.D. degrees from the Applied Mathematics Department, School of Mathematical Sciences, Peking University, Beijing, China, in 1999 and 2004, respectively. Currently, he is an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include computer vision, machine learning, and data mining.

THOMAS HUANG received the B.S. degree in electrical engineering from 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. From 1973 to 1980, he was with the School of Electrical Engineering, Purdue University, West Lafayette, IN, and Director of its Laboratory for Information and Signal Processing. In 1980, he joined the University of Illinois at Urbana-Champaign (UIUC), where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at the Coordinated Science Laboratory, Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology (UIUC), and Co-Chair of the Institute's major research theme: human-computer intelligent interaction. His interests are in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 20 books and many papers in network theory, digital filtering, image processing, and computer vision.

BINGFENG ZHOU was born in 1963, Ph.D., professor, Ph.D. supervisor. His research interests include graphics simulation of robot kinematics, geometry models and CAD/CAM, color image processing, multimedia system and image special effects, digital image halftone, image based rendering and modeling, virtual reality, etc.