Compressed Domain Real-time Action Recognition

Chuohao Yeo, Parvez Ahammad, Kannan Ramchandran and S Shankar Sastry
Department of Electrical and Computer Science
University of California, Berkeley
Berkeley, CA 94720–1770
Email: {zuohao,parvez,kannanr,sastry}@eecs.berkeley.edu

***Abstract*— We present a compressed domain scheme that is able to recognize and localize actions in real-time. The recognition problem is posed as performing a video query on a test video sequence. Our method is based on computing motion similarity using compressed domain features which can be extracted with low complexity. We introduce a novel motion correlation measure that takes into account differences in motion magnitudes. Our method is appearance invariant, requires no prior segmentation, alignment or stabilization, and is able to localize actions in both space and time. We evaluated our method on a large action video database consisting of 6 actions performed by 25 people under 3 different scenarios. Our classification results compare favorably with existing methods at only a fraction of their computational cost.**

I. Introduction

The use of video cameras has become increasingly common as their costs decrease. In personal applications, it is common for people to record and store personal videos that comprise various actions. In security applications, multiple video cameras record video data across a designated surveillance area. This proliferation of video data naturally leads to information overload. It would not only be incredibly helpful but also necessary to be able to perform rudimentary action recognition in order to assist users in focusing their attention on actions of interest.

In this paper, we would like to solve the following problem: given a query video sequence of a particular action, we would like to detect all occurrences of it in a test video, and thereby recognizing an action as taking place at some specific time and location in the video. We want our method to be appearance invariant, and we want a solution that can operate in real-time.

There has been prior work in action recognition using raw video without the use of body landmark points. Efros et. al. [1] require the extraction of a stabilized image sequence before using a rectified optical flow based normalized correlation measure for measuring similarity. Shechtman and Irani [2] exhaustively test motion-consistency between small space-time (ST) patches to compute a correlation measure between a query video and a test video. Schüldt et. al. [3] use a local feature based approach in which Support Vector Machines (SVM) were used to classify actions in a large database of action videos that they collected.

Any practical system that records and stores digital video is likely to employ video compression. It has long been recognized that some of the video processing for compression can be reused in video analysis or transcoding; this has been an area of active research (see for example [4], [5]) in the last decade or so.

There has also been prior work in performing action recognition in the compressed domain. Ozer et. al. [6] applied Principal Component Analysis (PCA) on motion vectors from *segmented* body parts for dimensionality reduction before classification. They require that the sequences must have a fixed number of frames and be temporally aligned. Babu et. al. [7] trained a Hidden Markov Model (HMM) to classify each action, where the emission is a codeword based on the histogram of motion vector components of the *whole* frame. In later work [8], they extracted Motion History Image (MHI) and Motion Flow History (MFH) [9] compressed domain features, before computing global measures for classification.

Our proposed method makes use of motion vector information to capture the salient features of actions which are appearance independent. It then computes frame-to-frame motion similarity with a measure that takes into account differences in both orientation and magnitude of motion vectors. The scores for each ST candidate are then aggregated over time using a method similar to [1]. Our approach is able to localize actions in space and time by checking all possible ST candidates, much like in [2], except that it is more computationally tractable since the search space is greatly reduced from the use of compressed domain features. Our novelty lies in our ability to perform *real-time* localization of actions in space and time by a novel combination of signal processing and computer vision techniques. The proposed method requires no prior segmentation, no temporal or spatial alignment and minimal training.

II. Proposed method

Given a query video template and a test video sequence, we carry out the following steps to compute a score for how confident we are that the action presented in the query video template is happening in each space-time location (to the nearest macroblock and frame) in the test video. We will elaborate on each of these steps in the following subsections.

1) Estimate optical flow from motion vectors.
2) Compute frame-to-frame motion similarity.
3) Aggregate similarities over a series of frames.
4) Repeat the above steps for all possible space-time locations to localize queried action.

A. Estimation of coarse optical flow

Motion compensation is an integral component of modern video compression technology and motion vectors are by-products of the process. Motion vectors are obtained from block matching and can be interpreted as crude approximations of the underlying motion field or optical flow. In addition, the DCT coefficients can also be used to provide a confidence measure on the estimate. We follow the approach outlined by Coimbra and Davies [10] for computing a coarse estimate and a confidence map of the optical flow.

We then threshold the confidence map to keep only the optical flow estimates with high confidence measures. This step removes unreliable estimates and greatly improves the performance of our algorithm.

B. Computation of frame-to-frame motion similarity

For the purpose of discussion in this section, both the test frame and query frame are assumed to have a spatial dimension of $N \times M$ macroblocks (the equal size restriction will be lifted later). We would like to measure the motion similarity between the motion field of the ith test frame, $\vec{V}_i^{\text{test}}(n,m)$, and that of jth query frame, $\vec{V}_j^{\text{query}}(n,m)$. We denote the horizontal and vertical components of a motion field $\vec{V}(n,m)$ by $V_x(n,m)$ and $V_y(n,m)$ respectively.

One way of measuring similarity is to follow the approach taken by Efros et. al. [1]. Each motion field is first split into non-negative motion channels, e.g. $(V_x(n,m))_+$, $(-V_x(n,m))_+$, $(V_y(n,m))_+$ and $(-V_y(n,m))_+$, where $(x)_+ = \max(0,x)$. We can then vectorize these channels and stack them into a single vector $\vec{U}$. The similarity between frame i of the test frame and frame j of the query frame, $S(i,j)$, is then computed as a normalized correlation:

$$S(i,j) = \frac{\langle \vec{U}_i^{\text{test}}, \vec{U}_j^{\text{query}} \rangle}{\|\vec{U}_i^{\text{test}}\| \|\vec{U}_j^{\text{query}}\|} \tag{1}$$

However, this does does not take into account the difference in magnitudes of the motion vectors. We propose a novel measure of similarity:

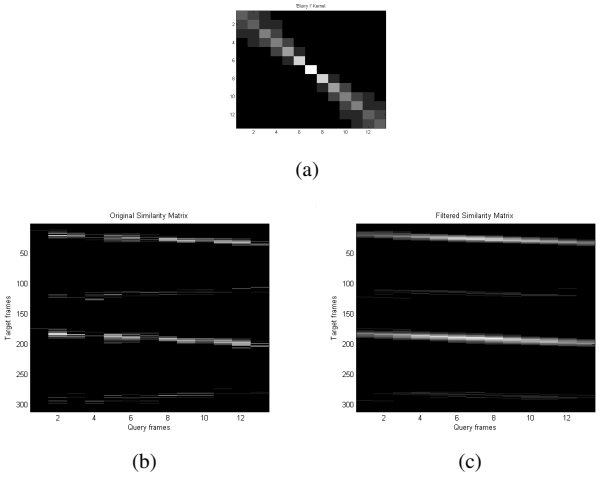
$$S(i,j) = \frac{1}{Z(i,j)} \sum_{n=1}^{N} \sum_{m=1}^{M} d(\vec{V}_i^{\text{test}}(n,m), \vec{V}_j^{\text{query}}(n,m)) \tag{2}$$

where if $\|\vec{V_1}\| > 0$ and $\|\vec{V_2}\| > 0$

$$\begin{aligned} d(\vec{V_1}, \vec{V_2}) &= \frac{\left(\langle \vec{V_1}, \vec{V_2} \rangle\right)_+}{\|\vec{V_1}\| \|\vec{V_2}\|} \cdot \min\left(\frac{\|\vec{V_1}\|}{\|\vec{V_2}\|}, \frac{\|\vec{V_2}\|}{\|\vec{V_1}\|}\right) \\ &= \frac{\left(\langle \vec{V_1}, \vec{V_2} \rangle\right)_+}{\max\left(\|\vec{V_2}\|^2, \|\vec{V_1}\|^2\right)} \end{aligned} \tag{3}$$

and $d(\vec{V_1}, \vec{V_2}) = 0$ otherwise. Also, the normalizing factor, $Z(i,j)$, in (2) is:

$$Z(i,j) = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbf{1}[\|\vec{V}_i^{\text{test}}(n,m)\| > 0 \text{ or } \|\vec{V}_j^{\text{query}}(n,m)\| > 0] \tag{4}$$



(a)

(b) (c)

Fig. 1. An example similarity matrix and the effects of applying aggregation. In these graphical representations, bright areas indicate a high value. (a) Aggregation kernel, (b) Similarity matrix before aggregation, (c) Similarity matrix after aggregation.

In other words, we want to ignore macroblocks in both the query and test video which agree on having no motion. This has the effect of not penalizing corresponding zero-motion regions in both the query and test video. We term this novel measure NZMS (Non-Zero Motion Block Similarity).

C. Aggregation of frame-to-frame similarities

Section II-B tells us how to compute $S(i,j)$. To take temporal dependencies into account, we need to perform an aggregation step. We do this by convolving the $S(i,j)$ with a $T \times T$ filter $H(i,j)$ to get an aggregated similarity matrix $S_a(i,j) = (S * H)(i,j)$ [1]. $S_a(i,j)$ tells us how similar a T-length sequence centered at frame i of the test video is to a T-length sequence centered at frame j of the query video. $H(i,j)$ can be interpreted as a bandpass filter that "passes" actions in the test video that occur at approximately the same rate as in the query video. We use the following filter [1]:

$$H(i,j) = \sum_{r \in R} \exp(-\alpha(r-1))\chi(i,rj) \ , \ -T/2 \le i,j \le T/2 \tag{5}$$

where

$$\chi(i,rj) = \begin{cases} 1 & \text{if } i = \text{sign}(rj) \cdot \lfloor |rj| \rfloor \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

R is the set of rates to allow for and α is a parameter ($\alpha \ge 1$) that allows us to control how tolerant we are to slight differences in rates; the higher α is, the less tolerant it is to changes in the rates of actions. Figure 1(a) shows this kernel graphically.

Figure 1(b) shows a pre-aggregation similarity matrix, $S(i,j)$. Note the presence of near-diagonal bands, which is a clear indication that the queried action is taking place in those frames. Figure 1(c) shows the post-aggregation similarity matrix, $S_a(i,j)$, which has much smoother diagonal bands.

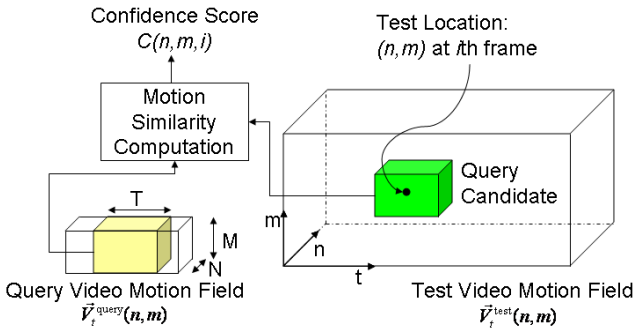


Fig. 2. Illustration of space-time localization. The query video space-time patch is shifted over the entire space-time volume of the input video, and the similarity, $C(n, m, i)$ is computed for each space-time location.

D. Space-time localization

Sections II-B and II-C tell us how to compute an aggregated similarity between each frame of a T_{test}-frames test sequence and each frame of a T_{query}-frames query sequence, both of which are $N \times M$ macroblocks in spatial dimensions. To compute an overall score on how confident we are that frame i of the test frame is from the query sequence, we use:

$$C(i) = \max_{\substack{\max(i-\frac{T}{2},1) \le k \le \min(i+\frac{T}{2},T_{\text{test}}) \\ 1 \le j \le T_{\text{query}}}} S_a(k, j) \tag{7}$$

Maximizing $S_a(k, j)$ over j of the query video allows us to pick up the best response that a particular frame of the test video has to the corresponding frame in the query video. We also maximize $S_a(k, j)$ over k in a T-length temporal window centered at i. The rationale is that if a T-length sequence centered at frame k of the test video matches well with the query video, then all frames in that T-length sequence should also have at least the same score.

The above steps can be easily extended to the case where the test video and query video do not have the same spatial dimensions. In that case, as proposed by Shechtman and Irani [2], we simply slide the query video template over all possible spatial-temporal locations (illustrated in figure 2), and compute a score for each space-time location using (7). This results in a action confidence volume, $C(n, m, i)$, which represents the score for the (n, m) location of the ith frame of the test video. A high value of $C(n, m, i)$ can then be interpreted as the query action being likely to be occurring at spatial location (n, m) in the ith frame.

While this exhaustive search seems to be computationally intensive, operating it in the compressed domain allows for a real-time implementation.

III. Experimental results

We evaluate our proposed algorithm on a comprehensive database compiled by Schüldt et. al. [3]. Their database captures 6 different actions (boxing, handclapping, handwaving, running, jogging and walking), performed by 25 people, over 4 different environments (outdoors, outdoors with scale variations, outdoors with different clothes and indoors). Within each environment, we compute the similarity, ρ, of each video (as the test video) to each of the other videos (as the query video) by first computing $C(n, m, i)$ as mentioned in section II-D, and then using:

$$\rho = \frac{1}{L} \sum_{i=1}^{T_{\text{test}}} \eta(i) \left(\max_{n,m} C(n, m, i) \right) \tag{8}$$

where the normalization factor L is given by:

$$L = \sum_{i=1}^{T_{\text{test}}} \eta(i) \tag{9}$$

and $\eta(i)$ is an indicator function which returns one if at least T frames in the $2T$-length temporal neighborhood centered at frame i have significant motion and returns zero otherwise. A frame is asserted to have significant motion if at least δ proportion of the macroblocks have reliable motion vectors of magnitude greater than ϵ. We then use leave-one-out K-nearest neighbor classification to label each of the videos.

In our experiments, we used $K = 9$, $\delta = \frac{1}{30}$, $\epsilon = 0.5$ pels/frame, $\alpha = 2.0$ and $T = 17$. For comparison, we also tested both normalized correlation (1) and NZMS (2). In addition, because our system does not handle scale-varying actions, we considered only the three environments that do not have significant scale variations.

A. Classification performance

The confusion matrix for NZMS is shown in table I, while that for normalized correlation [1] is shown in table II. Each entry of the matrix gives the fraction of videos of the action corresponding to its row that were classified as an action corresponding to the column. Our overall percentage of correct classification is 86%, which compares favorably to Schüldt et. al.'s [3] best reported result (just under 80%) on the same data set.

Looking at the confusion matrices, we see that our proposed NZMS measure vastly outperforms the simple normalized correlation measure. This is due to the fact that our measure looks at each corresponding pair of macroblocks separately instead of looking across all of them. NZMS also considers both differences in motion vector orientations and norms, and ignores matching zero-motion macroblocks.

Using NZMS, most of the confusion is between "Running" and "Jogging", with a significant proportion of "Jogging" videos being erroneously classified as "Running". Looking at the actual videos visually, we find it hard to distinguish between some "Running" and "Jogging" actions. There are cases where the speed of one subject's "Jogging" is faster than the speed of another subject's "Running"!!

B. Localization performance

Unlike most other methods, with the notable exception of [2], we are able to localize an action in space and time and as well as detect multiple and simultaneous occurring activities in the test video. Figure 3 shows an example (the "beach" test sequence and walking query sequence from Shechtman and

TABLE I

CONFUSION MATRIX USING NZMS

	Box	Hc	Hw	Run	Jog	Walk
Boxing	0.82	0.11	0.00	0.00	0.00	0.07
Hand**c**lapping	0.01	0.95	0.04	0.00	0.00	0.00
Hand**w**aving	0.07	0.04	0.89	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.91	0.00	0.09
Jogging	0.00	0.00	0.00	0.41	0.58	0.01
Walking	0.00	0.00	0.00	0.00	0.00	1.00

TABLE II

CONFUSION MATRIX USING NORMALIZED CORRELATION [1]

	Box	Hc	Hw	Run	Jog	Walk
Boxing	0.80	0.00	0.03	0.00	0.00	0.17
Hand**c**lapping	0.88	0.11	0.01	0.00	0.00	0.00
Hand**w**aving	0.09	0.00	0.87	0.00	0.00	0.00
Running	0.00	0.00	0.00	0.75	0.25	0.00
Jogging	0.01	0.00	0.00	0.01	0.98	0.00
Walking	0.00	0.00	0.00	0.55	0.00	0.45

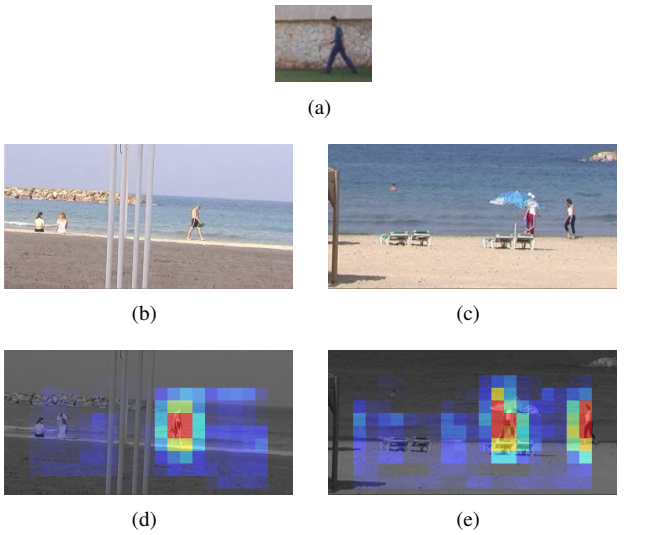


Fig. 3. Localization Results. The false color in (d) and (e) denotes detection responses, with blue and red indicating a low and high response respectively. (a) A frame from the query video, (b) An input video frame with one person walking, (c) An input video frame with two people walking, (d) Detection of one person walking, (e) Detection of two people walking.

Irani [2]) which demonstrates our algorithm's ability to detect multiple people walking in the test video.

C. Computational costs

On a Pentium-4 2.6 GHz machine with 1 GB of RAM, it took just under 11 seconds to process a test video of 368×184 pixels with 835 frames on a query video that is of 80×64 pixels with 23 frames. We extrapolated the timing reported in [2] to this case; it would have taken about 11 hours. If their multi-grid search was adopted, it would still have taken about 22 minutes. Our method is able to perform the localization, albeit with a coarser spatial resolution, up to *3 orders of magnitude faster*. On the database compiled in [3], each video has a spatial resolution of 160×120 pixels, and has an average of about 480 frames. For each environment, we would need to perform 22500 cross-comparisons. Yet, each run took an average of about 8 hours. In contrast, [2] would have taken an extrapolated run time of 3 years!

IV. CONCLUSION

We have designed, implemented and tested a system for performing action recognition and localization by making use of compressed domain features such as motion vectors and DCT coefficients which can be obtained with minimal decoding. The inherent reduction in search space makes real-time operation feasible. We combined existing tools in a novel way in the compressed domain for this purpose and also proposed NZMS, a novel frame-to-frame motion similarity measure. Our results compare favorably with existing techniques [3] on a publicly available database.

We plan to extend this to a multi-grid platform which would allow us to approach the spatial resolution of existing method at a lower computational cost. While our method is robust to small variations in scale, we would like to explore a truly scale-invariant approach in future work.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation (NSF) under Grant No. CCR-0330514. Chuohao Yeo is funded by the Agency for Science, Technology and Research, Singapore (A*STAR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or A*STAR.

REFERENCES

[1] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE International Conference on Computer Vision*, Nice, France, Oct. 2003.

[2] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, USA, June 2005, pp. 405–412.

[3] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach." in *Proc. International Conference on Pattern Recognition*, Cambridge, UK, Aug. 2004, pp. 32–36.

[4] S.-F. Chang, "Compressed-domain techniques for image/video indexing and manipulation," in *Proc. IEEE International Conference on Image Processing*, 1995, pp. 314–317.

[5] S. Wee, B. Shen, and J. Apostolopoulos, "Compressed-domain video processing," Hewlett-Packard, Tech. Rep. HPL-2002-282, 2002.

[6] B. Ozer, W. Wolf, and A. N. Akansu, "Human activity detection in MPEG sequences," in *Proc. IEEE Workshop on Human Motion*, Austin, USA, Dec. 2000, pp. 61–66.

[7] R. V. Babu, B. Anantharaman, K. Ramakrishnan, and S. Srinivasan, "Compressed domain action classification using HMM," *Pattern Recognition Letters*, vol. 23, no. 10, pp. 1203–1213, Aug. 2002.

[8] R. V. Babu and K. R. Ramakrishnan, "Compressed domain human motion recognition using motion history information," in *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003, pp. 321–324.

[9] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 928–934.

[10] M. T. Coimbra and M. Davies, "Approximating optical flow within the mpeg-2 compressed domain." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 103–107, 2005.