IMPLICIT MOTION-SHAPE MODEL: A GENERIC APPROACH FOR ACTION MATCHING

Tuan Hue Thi[1], *Li Cheng*[2], *Jian Zhang*[1], *Li Wang*[3]

[1]National ICT of Australia (NICTA) & University of New South Wales, 2032, NSW, Australia
[2]Toyota Technological Institute at Chicago, 60637, Illinois, USA
[3]Southeast University, Nanjing, 210096, China

ABSTRACT

We develop a robust technique to find similar matches of human actions in video. Given a query video, Motion History Images (MHI) are constructed for consecutive keyframes. This is followed by dividing the MHI into local *Motion-Shape* regions, which allows us to analyze the action as a set of sparse space-time patches in 3D. Inspired by the idea of Generalized Hough Transform, we develop the *Implicit Motion-Shape Model* that allows the integration of these local patches to describe the dynamic characteristics of the query action. In the same way we retrieve motion segments from video candidates, then project them onto the Hough Space built by the query model. This produces the matching score by running Parzen window density estimation under different scales. Empirical experiments on popular datasets demonstrate the efficiency of this approach, where highly accurate matches are returned within acceptable processing time.

Index Terms— Action Matching, Implicit Motion-Shape Model, Motion History Image, Generalized Hough Transform

1. INTRODUCTION

In this paper we consider the problem of finding similar matches of human actions, an important yet challenging task in automatic analysis and retrieval of video content.

Motion History Image (MHI), developed by Bradski and Davis [1], is an effective technique in combining the motion characteristics over consecutive time interval. In fact, many techniques have adopted this silhouette-based concept to extract the motion regions for each video, and use the obtained motion field as a 2D shape for recognition [1, 2, 3, 4, 8]. Despite the fact that these approaches are quite fast and simple, they are heavily dependent on how good the motion extraction, and are sensitive to motion noises.

Recent works [7, 9, 11, 12, 13] have focused on using local space time features to detect salient motion regions in video, and using scale-invariant local descriptors to capture the video content. The detected points are then fed into a trained classifier to infer the matching possibility of unknown data. These works rely on the intensive analysis of sparse regions, which focuses mainly on how to derive the characteristics of each common part of the action model and use them to differentiate between different actions, under different forms of integral optimization [11, 12]. In these approaches, the entity of each separate part is well defined, but the global structure is often neglected. This generally makes them vulnerable under circumstances when the core parts cannot be detected (occluded or weakly distinctive), or when actions having similar elements but totally different behaviors.

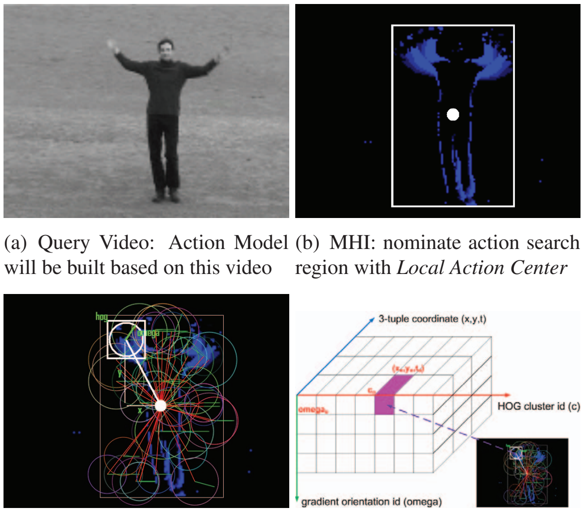
Our main contribution in this work is the development of a nonparametric model for action matching, which integrates local scale-invariant motion features (using Histogram of Oriented Gradients and clustered description on MHI) with global implicit structure of the action (implemented as the Hough Space projection). We developed a complete video search system based on the proposed matching algorithm, which performs competitively to the state-of-the-arts on standard testbeds such as KTH [12] and Weizmann [3] datasets.

2. MOTION-SHAPE FEATURES FOR VIDEO

Human action in our perspective is represented as a sparse set of local *Motion-Shape* features. Those are the selective patches detected at different scale from the Motion History Image (MHI), each contains information about the motion field and the shape of the actor, hence we loosely use the term *Motion-Shape* to describe. By analyzing these local attributes and their global configuration, we can conceptually articulate the behavior of the formulated action. Figure 1 describes our feature extraction algorithm. From the query video 1(a), we compute a collection of MHIs following [1]. For each MHI, there will be detected the dominant motion region and its center point (white rectangle and circle in Figure 1(b)). This motion blob and its center are used as references for *Motion-Shape* searching. In our design, we call those center points *Local Action Centers*, as opposed to *Global Action Center*, the mean position of all *Local Action Centers* in each video.

Sliding the window search at different scales in Figure 1(c) will help to produce a collection of *Motion-Shapes*, drawn as the thin colored circles. Each *Motion-Shape* m is

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications & Digital Economy and the Australian Research council through the ICT Centre of Excellence program



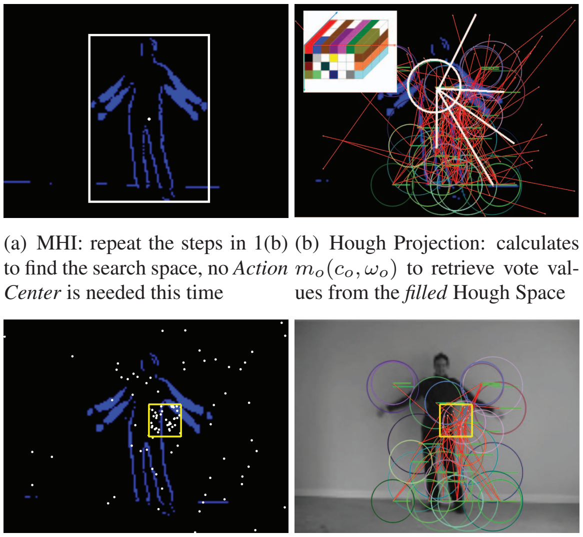
(a) Query Video: Action Model will be built based on this video

(b) MHI: nominate action search region with *Local Action Center*

(c) *Motion-Shape*: extracted as $m(x, y, t, c, \omega)$; circle color represents cluster identification

(d) Hough Space: represented as Lookup Table, 3-tuple $m(x, y, t)$ is filled using index pair $m(c, \omega)$

Fig. 1. Constructing Hough Space for Action Model



(a) MHI: repeat the steps in 1(b) to find the search space, no *Action Center* is needed this time

(b) Hough Projection: calculates $m_o(c_o, \omega_o)$ to retrieve vote values from the *filled* Hough Space

(c) Density Search: voted *Action Centers* as white points, *Fitting Region* is found in yellow square

(d) Action Segmentation: extract those *Motion-Shapes* contributing to the *Model Fitting Region*

Fig. 2. Fitting Video Candidate to Model

defined by its relative position (x, y, t) to the *Action Centers*. Since each *Motion-Shape* by its nature is a motion field, we easily compute the Histogram of Oriented Gradients (HOG), and also the orientation ω of the most dominant gradient in each patch. Using HOG values obtained from each patch, we cluster the patches into different representative group, each defined by a cluster identification number c. Figure 1(c) shows the cluster id of *Motion-Shape* circles using different colors. At this stage, a video shot is genuinely decomposed into a set of *Motion-Shapes* $m(x, y, t, c, \omega)$, which will be encoded into our model, described in the next section.

3. IMPLICIT MOTION-SHAPE MODEL

In order to produce a generic design for modeling the human action, we extend the Generalized Hough Transform [1] technique to detect the 3D action structure formed by different distinctive *Motion-Shapes*. Using *Global Action Center* as the reference point in our 3D structure, we build a Hough Space to quantitatively represent the relative position of all *Motion-Shapes* in the action model, illustrated as the Lookup Table in Figure 1(d). In that coordinate system, each *Motion-Shape* is indexed by a key pair $I = (c, \omega)$ consisting of its cluster id c and gradient orientation id ω. The $3-tuple$ entries (x, y, t) in the Lookup Table are filled by those *Motion-Shapes* attributes extracted from the model video based on index key value.

Providing this Hough Space, the action matching task now becomes projecting the *Motion-Shapes* collected from video candidate to this space, matching value will be calculated as how well those projected points fit in the model. Figure 2 illustrates the main steps of the matching task, starting with the MHI construction (Figure 2(a)) to calculate new *Motion-Shapes* in 2(b). Using the *filled* Model Hough Space (top left of 2(b)), a particular *Motion-Shape* m_o (circled in white) will use its index key (c_o, ω_o) to find its corresponding entries in the Lookup Table. Formally, with each hypothesis about *Action Center* position γ, we can factor the marginalization probability $p(\gamma|m, l)$ using *Motion-Shape* evidence e and its location $l = (x, y, t)$ based on the pair index $I = (c, \omega)$, this factorization is adopted from Leibe *et al.* [10]

$$p(\gamma|e, l) = \sum_i p(\gamma|e, I_i, l)p(I_i|e, l) \quad (1)$$

$$= \sum_i p(\gamma|I_i, l)p(I_i|e) \quad (2)$$

In our case, $p(I_i|e)$ is simply the calculation of (c, ω) from detected *Motion-Shapes*, and $p(\gamma|I_i, l)$ is the retrieval of *Action Center* location l using index key I_i. The quantitative matching score of an unknown motion compared to the action model is then defined as the summation of all *Motion-Shape* entities and *Action Center* locations

$$score(\gamma) = \sum_k \sum_j p(\gamma_j|e_k, l_k) \quad (3)$$

In our design, we run mean-shift Parzen window density estimation to find the *Model Fitting Region*, which is the projected region that has the highest density of voted Action Centers (Figure 2(c)). In our model, we adopt two kinds of reference points, *Global Action Center* and *Averaging Local Action Centers*. While the former runs 3D volume density searches using 3-tuple (x, y, t) location of a *Global Action*

Center, the latter runs 2D area searches using 2-tuple (x, y) location of all *Local Action Centers* on multiple MHIs and the average of all best matching will be used. We will discuss the effects of these two referencing systems in section 4. The search criteria for mean-shift is the ratio r of vote counts on searching region. In our video search system, we rank video relevancy based on this ratio in ascending order, best match has the highest r.

4. EXPERIMENTAL RESULTS

We developed a complete video search system based on the proposed action matching algorithm, Figure 4 shows a snapshot of its user interface. On the left panel, the Query Video is loaded and processed to generate the codebook dictionary of *Motion-Shapes* (in this particular query, there are 289 clusters obtained); also a Segmentation thumbnail is generated to illustrate the relative position of the nominated *Action Center*. The right panel shows the video from the database that have high matches with the query video. There are three different views for each match (Original, Motion-Shapes, or Segmentation). The results are ordered according to the matching score r and the relative matching correspondence is implied by the yellow square *Model Fitting Regions*.

In order to attain a thorough evaluation of our technique, we use this Video Search system to run the tests on the two datasets KTH [12] (2400 video shots of 6 actions) and Weizmann [3] (92 video shots of 10 actions). The common benchmark *train/test* amount for these two datasets is *2/3 Split* on KTH, 16 persons for training and 9 persons for testing, and *leave-one-out* on Weizmann, 8 persons for training and 1 person for testing. Since we are carrying out searching task, we only need one training sample per query. Therefore, in order to produce fair comparison with reported works, we do a random selection (with equal samples of each action) of the search queries, and run the search on the same testing amount.

In our evaluation process, we are also interested in understanding the accuracy-time tradeoff relationship and view change sensitivity. We conduct three independent test runs based on three principal methods, namely, *Global.Mirrored*, *Local.Mirrored*, and *Local.NonMirrored*. The first term indicates whether the *Global* or average of *Local Action Centers* is used as reference point, the second term specifies if the search is mirrored in left-right direction, that is, each video candidate will be flipped horizontally to generate a mirror of itself, the test is then done on both instances, and the maximum matching score is used. Empirical results show that on average, the *Global* search requires double the processing time of *Local* search, while the *Mirrored* algorithm is 1.5 times slower than *NonMirrored*. Figure 3 summarizes the performance of three techniques on KTH and Weizmann using the Receiver Operating Characteristic (ROC) curve.

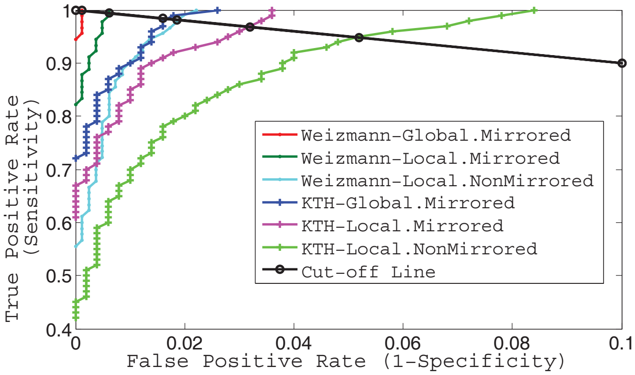
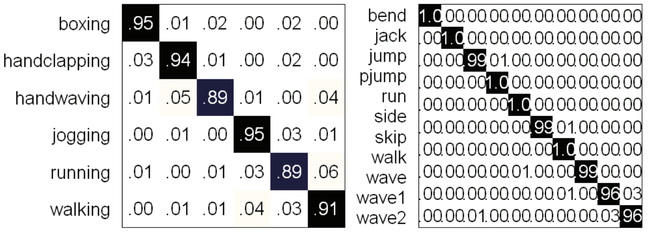


Fig. 3. Matching Performance on KTH and Weizmann

Tests on Weizmann dataset generally return better results than KTH, which is quite reasonable since the backgrounds in Weizmann are static, while the cameras in KTH are not stable. It is also shown that using *Global Action Center* as a reference for 3D point cloud search does yield better result than averaging individual 2D *Local Action Centers*. The performance decline in two *NonMirrored* techniques does imply that our technique is sensitive to view change, which is expected since we rely on the motion field shape to analyze action behavior, this drawback can be overcome with *Mirrored* method, sacrificing an overhead portion of processing time.

The black straight line in Figure 3 is called *Cut-off Line* which connects the two ends of True Positive Rate $[0, 1]$ and True Negative Rate $[1, 0]$, intersections of this line with ROC curves are called Cut-off points (represented as black circles), indicating the position where Sensitivity is equal to Specificity, and we use those points to analyze the average performance of our system specific for each action and as compared to other reported works on KTH and Weizmann. The performance of *Global.Mirrored* are elaborated as per action at *Cut-off points* using the Confusion Matrix with normalized sum on each row and column, as shown in Figure 5.



boxing	.95	.01	.02	.00	.02	.00
handclapping	.03	.94	.01	.00	.02	.00
handwaving	.01	.05	.89	.01	.00	.04
jogging	.00	.01	.00	.95	.03	.01
running	.01	.00	.01	.03	.89	.06
walking	.00	.01	.01	.04	.03	.91

(a) KTH 92.17%

bend	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
jack	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
jump	.00	.00	.99	.01	.00	.00	.00	.00	.00	.00
pjump	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
run	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
side	.00	.00	.00	.00	.00	.99	.01	.00	.00	.00
skip	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00
walk	.00	.00	.00	.00	.01	.00	.00	.99	.00	.00
wave1	.00	.00	.00	.00	.00	.00	.01	.00	.96	.03
wave2	.00	.00	.01	.00	.00	.00	.00	.00	.03	.96

(b) Weizmann 98.9%

Fig. 5. Matching Results based on Action

Using the average accuracy, we also conduct a comparison survey between our methods and reported *state-of-the-art* action recognition approaches, as shown in Table 4. Interestingly, while other works seem to work well for either one of the datasets, our system performs equally well, having the

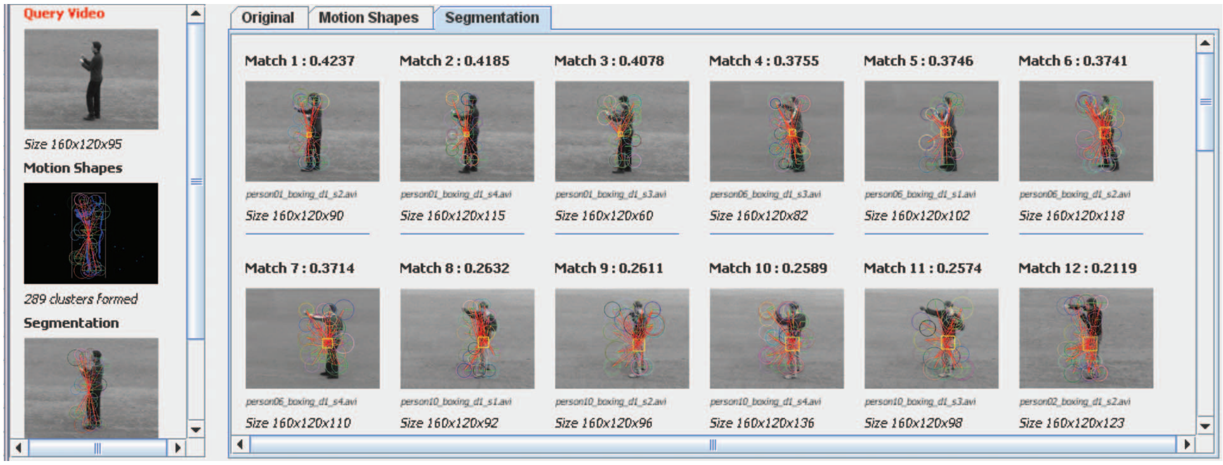


Fig. 4. Video Search GUI for Example using KTH *boxing* Action: Left Panel shows the loaded Query Video with nominated Action Centers. Right Panel displays the Best Matches returned, ranked using ratio r of vote counts in *Fitting Regions*

second best classification score on both KTH and Weizmann.

Methods	KTH	Weizmann
Global.Mirrored	***92.17***	***98.9***
Local.Mirrored	***85.93***	***94.75***
Local.NonMirrored	***79.2***	***85.12***
Schuldt *et al.* [12]	71.70	*
Ke *et al.* [8]	80.90	*
Niebles *et al.* [11]	81.50	72.8
Jiang *et al.* [6]	84.40	*
Fathi and Mori [2]	90.50	100
Jhuang *et al.* [5]	91.70	98.8
Wang and Mori [14]	91.17	98.33
Laptev *et al.* [9]	91.80	*
Grundmann *et al.* [4]	93.52	96.39

Table 1. Comparison with *state-of-the-art*

5. CONCLUSION AND FUTURE DIRECTION

We tackled the problem of action matching in videos using a nonparametric approach generalized from the Hough Transform algorithm. Dictating a query video by the its structured motion fields, we have successfully integrated local invariant motion features with global action configuration. Our system does not rely on any particular model parameter, hence makes it highly dynamic and generative. Ongoing work is improving the current system performance to realtime, producing a complete visual-based system for video search.

6. ACKNOWLEDGEMENT

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

7. REFERENCES

[1] G. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. *MVA*, 13:174–184, 2002.

[2] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, pages 1–8, 2008.

[3] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29:2247–2253, 2007.

[4] M. Grundmann, F. Meier, and I. Essa. 3d shape context and distance transform for action recognition. In *ICPR*, 2008.

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

[6] H. Jiang, M. Drew, and Z. Li. Successive convex matching for action detection. In *CVPR*, pages II: 1646–1653, 2006.

[7] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[8] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, pages 1–8, 2007.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic actions from movies. In *CVPR*, 2008.

[10] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an ism. In *ECCV*, 2004.

[11] J. Niebles, H. Wang, and F. Li. Unsupervised learning of human actions using spatial-temporal words. *IJCV*, 2008.

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

[13] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[14] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *PAMI*, 31(10), 2009.