Motion Divergence Fields for Dynamic Hand Gesture Recognition*

Xiaohui Shen[1] Gang Hua[2] Lance Williams[3] Ying Wu[1]

[1]Northwestern University
2145 Sheridan Road
Evanston, IL 60208

[2]IBM Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532

[3]Nokia Research Center Hollywood
2400 Broadway Street
Santa Monica, CA 90404

Abstract

Although it is in general difficult to track articulated hand motion, exemplar-based approaches provide a robust solution for hand gesture recognition. Presumably, a rich set of dynamic hand gestures are needed for a meaningful recognition system. How to build the visual representation for the motion patterns is the key for scalable recognition. We propose a novel representation based on the divergence map of the gestural motion field, which transforms motion patterns into spatial patterns. Given the motion divergence maps, we leverage modern image feature detectors to extract salient spatial patterns, such as Maximum Stable Extremal Regions (MSER). A local descriptor is extracted from each region to capture the local motion pattern. The descriptors from gesture exemplars are subsequently indexed using a pre-trained vocabulary tree. New gestures are then matched efficiently with the database gestures with a TF-IDF scheme. Our extensive experiments on a large hand gesture database with 10 categories and 1050 video samples validate the efficacy of the extracted motion patterns for gesture recognition. The proposed approach achieves an overall recognition rate of 97.62%, while the average recognition time is only 34.53 ms.

1. Introduction

Vision-based hand gesture recognition has been an active research topic for decades due to its wide applications in human computer interfaces, robot control, and mixed/augmented reality, *etc.*. It can be conceptually divided into static hand gesture recognition, and dynamic hand gesture recognition. A comprehensive overview of recent gesture recognition methods can be found in [14]. Due to the motion information, dynamic hand gestures offer a rich communication channel.

The approaches to dynamic hand gesture recognition can be categorized into two: model-based, and exemplar-based. For the model-based approach, Hidden Markov Models (HMM) are perhaps most frequently used [22, 19, 5, 21]. In [22], dynamic feature vectors are transformed to symbolic sequences by vector quantization, and subsequently modeled by a discrete HMM. Some recent improvements over traditional HMM include the semantic network model (SNM) [19], the non-parametric HMM [5], and the Hidden Conditional Random Field [21]. These variants either reduce training efforts, or improve classification accuracy.

Other model-based approaches include Finite State Machines (FSM) [3, 10], dynamic Bayesian Networks (DBN) [20], and topology-preserving self-organizing networks [7]. All these approaches assume that the hand has been detected and its articulated motion is tracked, which is either achieved by skin color segmentation, or kinematic model based hand tracking [1]. Although they have delivered promising results, the robustness of these approaches is dependent on the prior success of (frequently challenging) hand detection and motion tracking. Furthermore, it is both data intensive and computationally difficult to train these models before they can be applied in recognition.

To circumvent the difficulties of hand tracking and detection, other approaches try to leverage invariant visual representation for matching and recognition. Among these descriptors, local spatio-temporal features are the most widely exploited [9, 18, 4]. Freeman and Roth [9] propose that spatio-temporal gradients may be useful for dynamic gesture recognition. It is further validated by [18] in which they use SVM to classify space-time interest points for action recognition. Dollár *et al.* [4] extract histograms of 3D spatio-temporal cuboids to match human actions. Other visual features include motion trajectories [23] and Gaussian Density Features surrounding selected interest points [11]. More recently, Chaudhry *et al.* [2] calculate a sequence of histograms of oriented optical flow and use Binet-Cauchy kernels on nonlinear dynamical systems to achieve state-of-the-art results in periodic human action recognition.

However, none of these methods offer a scalable solu-

*The majority of the work was conducted when Xiaohui Shen was a summer intern supervised by Gang Hua at Nokia Research Center Hollywood.

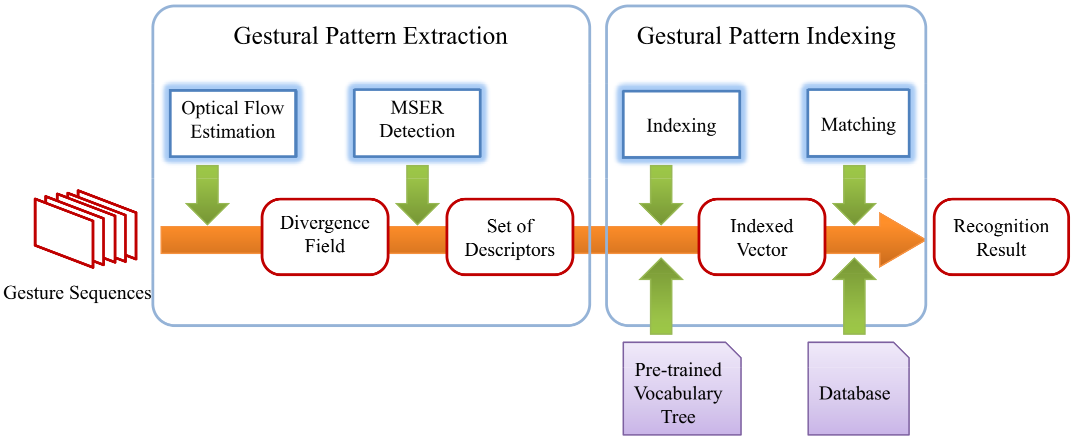


Figure 1. The pipeline of the proposed method.

tion for efficient matching when the exemplar database is large. In previous work, to improve the run-time speed for gesture recognition, most efforts have been devoted to speeding up image and video processing by either designing efficient algorithms [17] or leveraging more powerful computing units [6] such as a GPU. This is certainly desirable. Nevertheless, the issue of building efficient visual representation for scalable hand gesture matching over large exemplar databases has not been adequately addressed.

We propose a new visual representation of dynamic hand gestures based on the divergence field of the flow motions, which transforms gestural motion patterns into spatial image patterns. Given a gesture sequence, we extract the optical flow between any two consecutive frames. Its divergence field is derived. We then detect salient spatial patterns from the divergence field using a Maximally Stable Extremal Regions(MSER) [13, 8] feature detector. A descriptor is subsequently extracted from each detected region to characterize the local motion patterns. The descriptors extracted from all example gesture sequences are indexed by a pre-trained hierarchical vocabulary tree. A new gesture sequence is recognized by matching against the database with a TF-IDF scheme [15], which is scalable to large databases. The pipeline of our method is illustrated in Fig.1.

To the best of our knowledge, this paper is a first attempt to transform local motion patterns into spatial image patterns using the divergence field. This enables us to leverage state-of-the-art image indexing techniques for scalable hand gesture recognition without resorting to hand detection, segmentation, or tracking. We have collected a sizable database of dynamic hand gestures with 10 categories and 1050 samples for evaluation, which will be shared with the research community for further study. The recognition rate of our method on our evaluation dataset is 97.62%, with average recognition time 34.53ms. In other words, our proposed approach presents not only a novel approach to motion pattern analysis, but also a scalable framework for dynamic hand gesture recognition with a large number of examples.

2. Visual patterns of gestural motion

In this section, we present the proposed visual representation of gestural motions. In essence, optical flow is estimated between every two consecutive frames and the motion divergence field is extracted, from which MSER regions and local motion descriptors are then extracted.

2.1. The divergence field of optical flow

In a vector field, divergence is an operator that measures the magnitude of the source or sink of the field. Given a vector $\mathbf{F} = [F_1, F_2, \cdots, F_n]^T$ in a n-dimensional Euclidean space, the divergence of $\mathbf{F}$ can be calculated as:

$$\mathbf{divF} = \sum_{i=1}^{n} \frac{\partial F_i}{\partial x_i}, \tag{1}$$

where $[x_1, x_2, \cdots, x_n]^T$ are the Cartesian coordinates of the space where the vector field is defined.

Accordingly for an optical flow vector field $\mathbf{F}(x, y) = [u(x, y), v(x, y)]^T$, where $u(x, y)$ and $v(x, y)$ are respectively the horizontal and vertical components of optical flow at position (x, y), the divergence of $\mathbf{F}$ is:

$$\mathbf{divF} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}. \tag{2}$$

Fig.2 presents an example of transforming a flow motion field into a divergence field. Fig.2(a) is the first frame of an image pair, and Fig.2(b) and (c) are the visualization of u and v respectively. The corresponding divergence field is shown in Fig.2(d). We calculate the optical flow using the Lucas-Kanade algorithm[12] and implement the algorithm on the GPU to speed up the processing frame rate. The

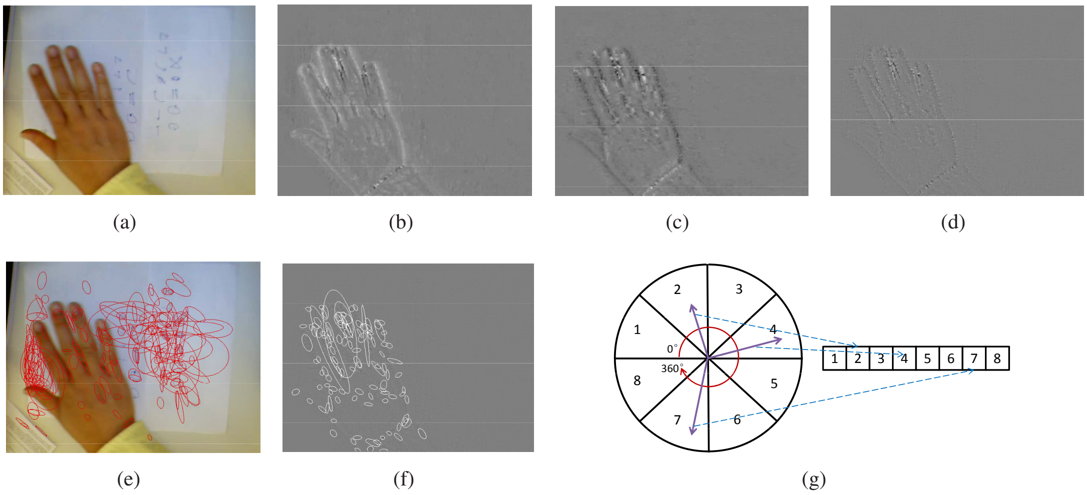


Figure 2. Gestural pattern extraction.(a) the first frame of an image pair, (b) the u component of the estimated optical flow, (c) the v component of the estimated flow, (d) the divergence field, (e) MSER detection directly on the image in (a), (f) MSER detection on the divergence field in (d), (g) calculating a histogram of optical flow orientations with 8 bins from MSER regions.

Lucas-Kanade method, as applied here, directly estimates flow at each pixel. Such estimation relies on local contrast and texture, and is valid only for small motions. Since no multiresolution estimation is applied, the computed flow is not valid for large areas with little texture, such as the interior of the hand region in the example. As a result, the flows in these areas can hardly be estimated. However, such estimate of motion proved a sound basis for discrimination in our experiments, even if the flow values are not absolutely accurate. As we can see, the flow divergence field has filtered out most flow noise in the background and provides a clear shape of the hand, which ensures that most MSER regions are located on the hand. We proceed to present the extraction of the local motion descriptors.

2.2. Local descriptor extraction

Once we obtain the divergence field, we perform Maximally Stable Extremal Regions (MSER) [13] detection. The MSER region is defined exclusively by an extremal property of the intensity function in the region and on its outer boundary, and therefore has many useful properties, including invariance to affine transformation of image intensities, stability, and allowing multi-scale detection. MSER has been widely used in image matching and object recognition, and has led to better recognition performance [8] in several applications. In our proposed framework, each MSER region is fitted by an ellipse, as shown in Fig.2(f). Because the background is static, most MSER regions detected in the motion divergence field are on the boundary of the hand or within the hand. The features extracted from these regions therefore are not mixed with background clutter. As a comparison, Fig.2(e) shows the MSER regions derived directly from the image in Fig.2(a), which are as frequently detected in background regions as within the moving hand.

In each detected MSER region, we calculate a histogram of the orientations of the optical flow vectors. The orientations of optical flow can be calculated from $u(x, y)$ and $v(x, y)$ and have a range of $[0, 2\pi]$. All the orientations are then bi-linearly quantized and aggregated into discrete bins with their magnitudes as weights. Fig.2(g) gives us a simple illustration, in which the histogram has 8 bins. In practice we set the bin number to be 80. The histogram is finally normalized to have unit L1-norm.

The rationale for choosing histograms of flow orientations as our local descriptors is that the speed of gestures varies widely, particularly among different users. Hence a good gesture recognition algorithm should be relatively insensitive to the speed with which a gesture is performed. This suggests orientations of hand movement as significant measures for recognition. What we are seeking here is a set of discriminative descriptors for each distinct gesture. We validate in our experiments that such descriptors are already highly discriminative, irrespective of the fact that we adopted a simple algorithm to estimate optical flow.

After local descriptor extraction, each divergence field is represented by a set of local descriptors, and a hand gesture is a sequence of such descriptor sets. We dispense with motion estimation techniques such as segmentation and tracking. MSER detection can be performed at modest computational expense; we apply a recent linear time implementation [16]. The whole feature extraction process is very efficient, and high frame-rate performance is readily achieved.

3. Motion gestural pattern indexing

In this section, we will focus on a scalable indexing scheme based on the visual representation of the motion gestural patterns proposed in Section 2 for dynamic hand gesture recognition with a large number of exemplars.

3.1. Vocabulary tree

In the context of image search and object recognition, a vocabulary tree is a hierarchical structure of cluster centers of a set of training descriptors. It was first used in [15] for object recognition. Fig.3 shows a simple vocabulary tree with 6 branches and 3 levels.

The vocabulary tree can be built by hierarchical k-means clustering. Given the training dataset, a k-means clustering process is first performed to determine k cluster centers, where k is the branch factor of the tree, i.e., the number of children of each node. In Fig.3, $k = 6$. These k centers represent the nodes in the first level of the tree. The training descriptors are then branched to k groups according to their distances from the cluster centers. In each group, k-means clustering is further performed to define k new cluster centers, which are then the children of the original center. The same process is carried out recursively until the tree achieves a pre-defined maximum depth. For a vocabulary tree with k branches and l levels, the total number of leaf nodes would be k^l.

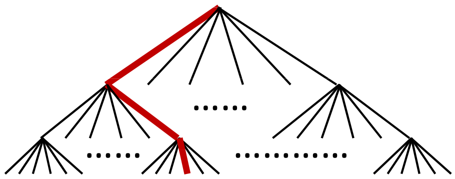


Figure 3. A vocabulary tree with 6 branches and 3 levels.

Once the vocabulary tree is built, the descriptors in an image can be quantized by comparing with the descriptors of the k nodes at each level, and associated with the closest one. Each descriptor thus has a path from the root to a leaf in the tree. The red line in Fig.3, for example, is a path for one descriptor. Such a path can be encoded by a single integer at the leaf, and used for indexing and matching, as described in the following section. Note that each quantization process for a descriptor involves only $k*l$ comparisons. The computational cost is logarithmic in the number of leaf nodes, which is the principal advantage offered by hierarchical structure.

3.2. Indexing a single image

After all the descriptors of a query image are quantized through the vocabulary tree, the image can be matched with the database images by comparing the similarities of the paths of their descriptors. Consider that n_i and m_i are the number of descriptors quantized to the i-th node in the query image and in a database image respectively, the distance between the query image and the database image can be defined as:

$$a_i = n_i w_i \,, \qquad b_i = m_i w_i$$

$$\begin{aligned} d(\mathbf{a},\mathbf{b}) &= \left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\|_p^p \\ &= \sum_i |a_i - b_i|^p \end{aligned} \tag{3}$$

where w_i is the weight of the i-th node in the vocabulary tree; p indicates Lp-norm. We use L1-norm in our experiments. The weight w_i can be defined based on entropy:

$$w_i = log \frac{N}{N_i} \tag{4}$$

where N is the number of images in the database, and N_i is the number of images that have descriptors quantized to the i-th node, which in text analysis is called *inverse document frequency*. It is found in practice that leaf nodes contain the most information, and sometimes only the leaf nodes are used in image matching for convenience.

Usually there are thousands of leaf nodes in a vocabulary tree, with only hundreds or dozens of descriptors in an image. As a result, **a** and **b** are both sparse vectors. After we normalize **a** and **b** to have unit magnitude, the distance of **a** and **b** in Eqn.3 can be further rewritten as:

$$\begin{aligned} d(\mathbf{a},\mathbf{b}) &= \sum_i |a_i - b_i|^p \\ &= \sum_{b_i=0} |a_i|^p + \sum_{a_i=0} |b_i|^p + \sum_{a_i\neq0,b_i\neq0} |a_i - b_i|^p \\ &= \sum_i |a_i|^p + \sum_i |b_i|^p \\ &\quad - \sum_{a_i\neq0,b_i\neq0} (|a_i - b_i|^p + |a_i|^p + |b_i|^p) \\ &= 2 - \sum_{a_i\neq0,b_i\neq0} (|a_i - b_i|^p + |a_i|^p + |b_i|^p) \end{aligned} \tag{5}$$

Only the distances between the non-zero elements of the vectors are calculated. This allows us to use inverted files to avoid direct matching of the query vector with all the image vectors in the database. An inverted file for each node is a file recording the number of images that have at least one descriptor quantized to that node, and the IDs of these images, along with the number of descriptors in these images that are quantized to that node, which in text analysis is called *term frequency*. Our distance measure incorporates a TF-IDF (*term frequency-inverse document frequency*) weighting scheme. When the descriptors of the query image are all quantized, only the inverted files of the nodes corresponding to the non-zero elements of the query vector are looked up. The distances of the query image to each of the images recorded in the inverted files can be gradually accumulated using Eqn.5. By using inverted files for matching in the TF-IDF scheme, the computational cost of image matching is significantly reduced. This approach enables efficient search even if there are millions of leaf nodes and database images.

3.3. Indexing a gestural motion sequence

The image indexing technique introduced in Section 3.2 only works in single image search, while in Section 2 a hand gesture is converted to a variable-length sequence of divergence images. In this section, we will extend image indexing for gesture sequence matching.

One straightforward solution is to uniformly sample frames from a gesture sequence. Using the method in Section 3.2, each sampled frame can be indexed to form a vector $\mathbf{a}^i = [a_1^i, a_2^i, \cdots, a_M^i]^T$, where i indicates the i-th sampled frame, and M is the number of leaf nodes in the vocabulary tree. We concatenate the vectors of all the sampled frames to form a new vector, which represents the indexing results of the entire gesture sequence. The distance of two gesture sequences accordingly can be calculated as:

$$d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} d(\mathbf{a}^i, \mathbf{b}^i) = \sum_{i=1}^{n} \sum_{j} \left| a_j^i - b_j^i \right|^p \qquad (6)$$

where n is the total number of sampled frames.

This extension, though straightforward and simple, is very important in our method for the following two reasons:

1. It normalizes the gestures to vectors with the same length, which removes the factor of gesture duration, and enables recognition of gestures with substantial variation in speed;

2. The direction of optical flow changes continually in some gestures (e.g., drawing a circle). Experimental results have validated that by sampling above a critical rate, the dynamic changes of the motion patterns can be successfully preserved in the concatenated vectors.

After the indexing procedure, the top k candidate gestures are returned. The positions of the MSER regions are then used to estimate a geometric center of the hand for each sampled frame. The centers in the query and those in the candidates are compared to assign a score to each candidate. This is a natural post-verification step in image search. The aggregated scores for each gesture category can be obtained from these top k scores, and the query gesture is then assigned to the category with the highest score.

4. Experiments

4.1. The database

The gestures in our database are 2D hand movements on an arbitrary background, with a static camera directly above the hand. It simulates scenarios in which users make hand gestures in front of a camera sitting on a tabletop, which can be applied in human interactions with mobile devices. There are 10 dynamic hand gestures in total, including: move right, move left, rotate up, rotate down, move down-right, move right-down, clockwise circle, counter-clockwise circle, “Z”, and “cross”, as shown in Fig.4(b). In the collection process, each person is asked to perform these ten actions with seven postures as illustrated in Fig.4(a): thumb, index finger, hand - fingers extended, “okay”(thumb and forefinger loop), fist, index finger with 90° rotation and hand with extended fingers at 90° rotation. Each subject contributes 70 gesture samples to our database. We collected 1050 sample gestures performed by 15 subjects. Fig.4(c) provides some example sequences. As we can see, the background as well as the skin colors of the hands are very diverse, and the captured sequences contain severe motion blurs. Both conditions are common in real applications. We consider this a representative database that is useful not only for dynamic hand gesture recognition but also for static hand pose estimation as well. We plan to share the dataset with the research community in the near future.

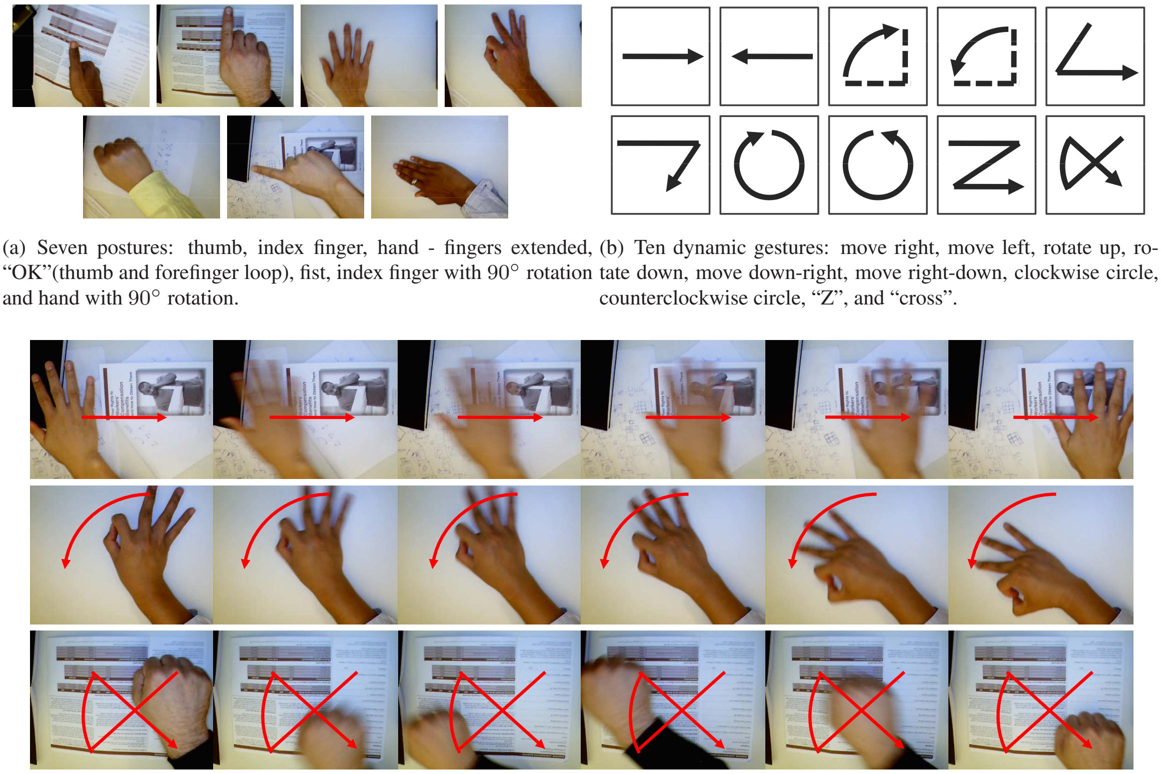
Since our method addresses dynamic hand gesture recognition, we focus on recognizing the 10 dynamic gestures in our experiments. Thus the samples with the same action in different hand postures are considered as one category, and each category accordingly has 105 samples.

4.2. Determining the parameters

There are two main parameters in our framework to be determined: the size of the vocabulary tree (the number of leaf nodes), and the number of sampled frames in a gesture sequence for indexing, as described in Section 3.3.

We have collected 2257851 descriptors in total, and chose different branch factors and levels to build the vocabulary tree. The recognition performance with different tree sizes is shown in Fig.5. Contradicting the observation in [15] that a larger vocabulary tree would improve performance, in our experiments the recognition rate is already very high with only 512 leaf nodes. The performance remains stable when the size is smaller than 2000, and drops slightly when the size is larger. However, the recognition rate is still above 96% when the size is over 10000. The vocabulary tree with 9 branches and 3 levels (729 leaf nodes) achieves the highest recognition rate, which is 97.62%. This tree will be used in the following experiments.

We tried encoding sequences in different numbers of sampled frames, and depict the results in Fig.6. As shown by the red line in Fig.6, recognition performance is quite robust at different frame sampling rates. Even if only 3 frames are sampled, the recognition rate has already achieved nearly 96%. When the frame sampling rate is larger than 7, the performance cannot be further improved.



(a) Seven postures: thumb, index finger, hand - fingers extended, "OK"(thumb and forefinger loop), fist, index finger with 90° rotation and hand with 90° rotation.

(b) Ten dynamic gestures: move right, move left, rotate up, rotate down, move down-right, move right-down, clockwise circle, counterclockwise circle, "Z", and "cross".

(c) Some examples in the database.

Figure 4. The collected database, with 10 categories of gestures and 1050 samples in total.

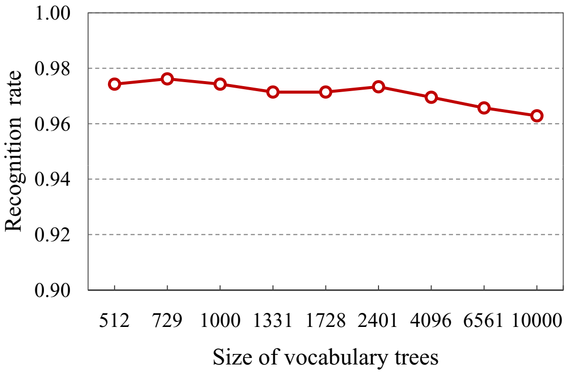


Figure 5. Recognition results with different vocabulary tree sizes.

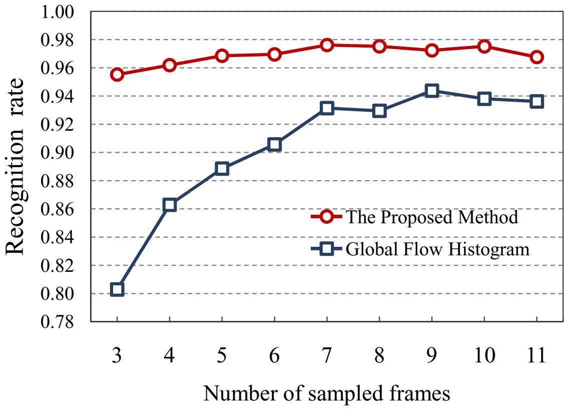


Figure 6. Recognition results with different sampling rates.

That is probably because, for discriminating gestures in our database, dynamic information of the motion patterns has already been fully captured in 7 frames.

4.3. Comparisons

We compared our method with two other methods. Leave-one-out cross-validation is used in the evaluation for all three methods. The first method is a baseline method in which MSER detection is directly performed on the image sequences, without optical flow estimation. HOG-like image descriptors are then sampled and indexed using the method in Section 3. Compared with our method which indexes and matches motion patterns from the divergence field, this baseline method directly matches appearance patterns of the hand. Variations in appearance would seem to necessitate a very large training set, and matching is greatly influenced by features extracted from the background. The performance of this method is poor in our dataset, barely above 50%, as shown in Fig.7 and Fig.8(a).

The second method is one adapted from Chaudhry *et al.* [2]. In their original paper, Chaudhry *et al.* [2] use a histogram for the entire oriented optical flow field and then extract the dynamics of the histograms for periodic action recognition. However, since only the dynamics (changes) in the histograms are used, this approach cannot be expected to discriminate some gestures in our database (e.g. constant motions left or right, in which the histograms do not necessarily change). To make their method work in our scenarios for a fair comparison, once we extract a sequence of histograms, we also resample frames to normalize sequence length, as proposed in Section 3.3, and calculate the χ^2 distance for the concatenated histograms:

$$d(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \sum_i \frac{|a_i - b_i|^2}{a_i + b_i} \tag{7}$$

The recognition result is then determined by a k-Nearest Neighbor classifier (k-NN). The best recognition rate of this method is 94.38%, which validates the discriminative power of histograms of oriented optical flow. However, our method is more robust than the global histogram. Fig.7 shows the recognition results of the three methods with different numbers of top k examples for classification. The recognition rate of our method is always higher than the global histogram. Moreover, the global histogram method is more sensitive to the sampling rate. As shown in Fig.6, The classification performance of global histograms drops dramatically when the sampling rate is under 7. This strongly suggests that the local descriptors for motion patterns in our method are more robust than global flow histograms in capturing and discriminating the dynamics of hand gestures.

Meanwhile, benefiting from the indexing scheme, the average recognition time of our method is 34.53 ms, only about 20% of the recognition time of global histogram matching, as shown in Fig.8(b). The recognition time of histogram matching is linear in the number of gestures in the database. Given a database with more than 10 thousand gestures, real-time recognition by nearest-neighbor matching is not currently feasible, while our method is readily scalable to large databases. Direct image indexing also takes more recognition time than our method, which is probably because many more MSER regions are detected directly from the images than from their optical flow divergence fields.

Fig.9 shows the confusion matrix of the recognition result of our method. The misclassification of some samples are due to the similarities of some gestures as well as the ambiguity when users perform these gestures. For example, in Fig.10, the user is doing the *Rotate Up* gesture. However, the rotation of the hand is not sufficient, and our method misclassifies the gesture as *Move Right*.

4.4. Applications

We built a prototype system and applied our method for live hand gesture recognition. Fig.11 shows a snapshot of this system. Gesture spotting, i.e., automatic detection of the start and end of the gestures, is implemented in the sys-

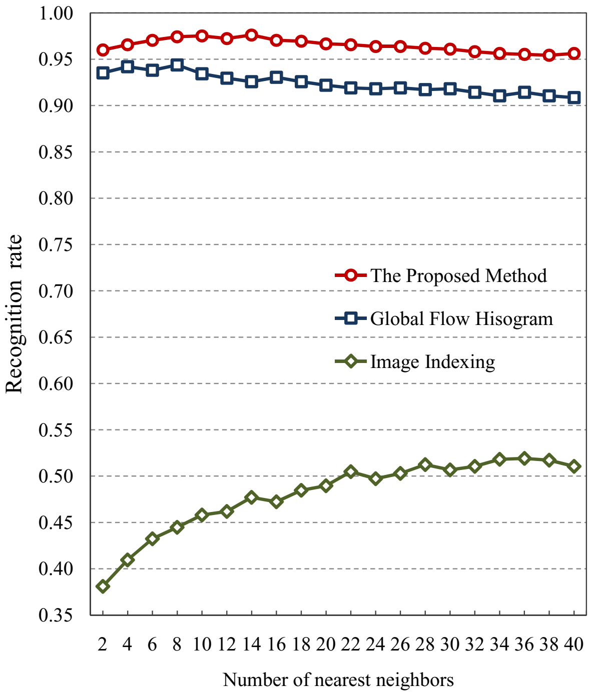
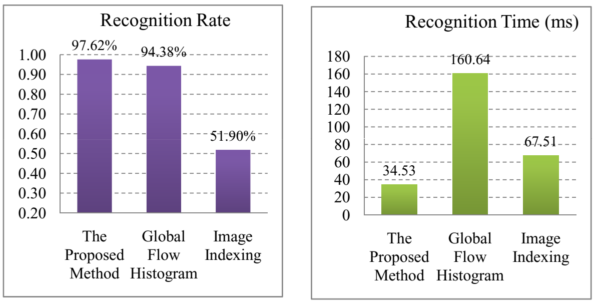
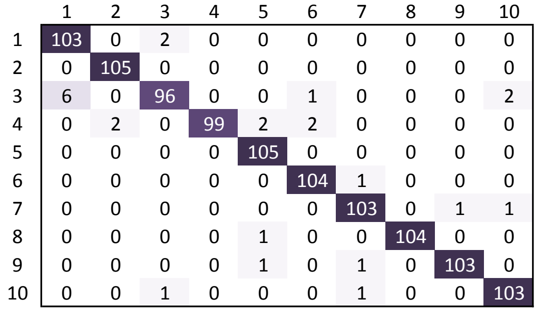


Figure 7. Performance with different top k candidates.



(a) (b)

Figure 8. Comparisons of three methods on recognition performance and time. Our method achieves the best performance within the shortest time.



	1	2	3	4	5	6	7	8	9	10
1	103	0	2	0	0	0	0	0	0	0
2	0	105	0	0	0	0	0	0	0	0
3	6	0	96	0	0	1	0	0	0	2
4	0	2	0	99	2	2	0	0	0	0
5	0	0	0	0	105	0	0	0	0	0
6	0	0	0	0	0	104	1	0	0	0
7	0	0	0	0	0	0	103	0	1	1
8	0	0	0	0	1	0	0	104	0	0
9	0	0	0	0	1	0	1	0	103	0
10	0	0	1	0	0	0	1	0	0	103

Figure 9. Confusion matrix of our recognition result. Each category contains 105 samples. The overall recognition rate is 97.62%.

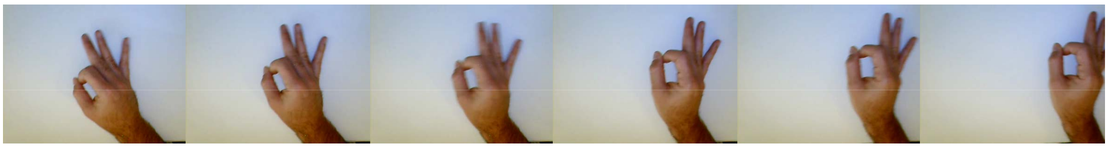


Figure 10. An example in which our method fails to recognize the gesture. In a *Rotate Up* gesture, the rotation of the hand is not great enough to be detected, and our method misclassifies the gesture as *Move Right*.

tem by a global motion threshold. The system can recognize live hand gestures in real-time (30 *fps*) for 320×240 video sequences, with high recognition quality.

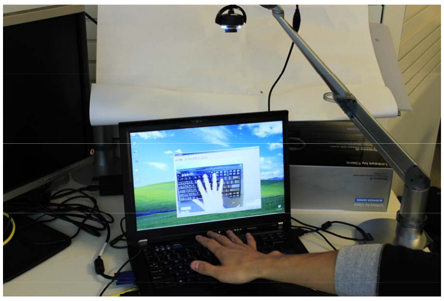


Figure 11. A snapshot of our prototype system

5. Conclusions

We presented a new method for dynamic hand gesture recognition, in which a gesture sequence is converted to a sequence of motion divergence fields. Deriving the optical flow divergence field transforms motion patterns into discriminative spatial image patterns, which are then extracted using a MSER detector, and indexed by a trained vocabulary tree. A TF-IDF scheme is utilized to match a query gesture to the indexed database sequences and to generate the final recognition result. The proposed framework is scalable to a large database, and simultaneously achieves high recognition accuracy. We believe that the proposed approach is also applicable to more general action/activity recognition tasks, and merits further study.

References

[1] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *FG*, pages 423–428, 2002.

[2] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, pages 1932–1939, 2009.

[3] J. Davis and M. Shah. Recognizing hand gestures. In *ECCV*, 1994.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS*, 2005.

[5] A. Elgammal, V. Shet, Y. Yacoob, and L. Davis. Learning dynamics for exemplar-based gesture recognition. In *CVPR*, 2003.

[6] C. L. H. F. S. Chen, C. M. Fu. Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Vision Computing*, 21:745–758, 2003.

[7] F. Flórez, J. M. García, J. García, and A. Hernández. Hand gesture recognition following the dynamics of a topology-preserving network. In *FG*, 2002.

[8] P. Forssen and D. Lowe. Shape descriptors for maximally stable extremal regions. In *ICCV*, 2007.

[9] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *FG*, 1995.

[10] P. Hong, M. Turk, and T. S. Huang. Gesture modeling and recognition using finite state machines. In *FG*, 2000.

[11] T. Kirishima, K. Sato, and K. Chihara. Real-time gesture recognition by learning and selective control of visual interest points. *IEEE Trans. on PAMI*, 27:351–364, 2005.

[12] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[14] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Trans. on Systems, Man and Cybernetics - Part C*, 37(3):311–324, 2007.

[15] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[16] D. Nistér and H. Stewénius. Linear time maximally stable extremal regions. In *ECCV*, pages 183–196, 2008.

[17] R.Lockton and A. W. Fitzgibbon. Real-time gesture recognition using deterministic boosting. In *BMVC*, 2002.

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

[19] S.Rajko, G. Qian, T. Ingalls, and J. James. Real-time gesture recognition with minimal training requirements and on-line learning. In *CVPR*, 2007.

[20] H. Suk, B. Sin, and S. Lee. Recognizing hand gestures using dynamic bayesian network. In *FG*, 2008.

[21] S. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, 2006.

[22] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, 1992.

[23] M. H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Trans. on PAMI*, 24:1061–1074, 2002.