Digit Recognition using Optical Flow Approach

Dr. Shallu Bashambu[1] Vishal Gupta[2] Mayank Sandilya[3]
[1]Associate Professor
[1,2,3]Department of Information Technology
[1,2,3]Maharaja Agrasen Institute of Technology Sector-22, Rohini, New Delhi-110086, India

Abstract— In this study, a new model to the problem of video action recognition has been proposed. The model is based on temporal video representation for automatic annotation of videos. Video action recognition is a field of multimedia research enabling us to recognize the actions from a number of observations, where representation of temporal information becomes important. Visual, audio and textual features are important sources for representation. Although textual and audio features provide high level semantics, retrieval performance using these features highly depends on the availability and richness of the resources. Visual features such as edges, corners, interest points etc. are used for forming a more complicated feature, namely, optical flow. For developing methods to cope with video action recognition, we need temporally represented video information. For this reason, we propose a new temporal segment representation to formalize the video scenes as temporal information. The representation is fundamentally based on the optical flow vectors calculated for the frequently selected frames of the video scene. Weighted frame velocity concept is put forward for a whole video scene together with the set of optical flow vectors. The combined representation is used in the action based video segment classification. Proposed method is applied to significant data sets and the results are analyzed by comparing to the state- of-the art methods.

Key words: Video Action Recognition, Content-Based Video Information Retrieval, Optical Flow

I. INTRODUCTION

Video action recognition is a field of multimedia research enabling us to recognize the actions from a number of observations. The observations on video frames depend on the video features derived from different sources. While textual features include high level semantic information, they cannot be automated. The recognition strongly depends on the textual sources which are commonly created manually. On the other hand, audio features are restricted to a supervisor role. As the audio does not contain strong information showing the actions conceptually, it can be used as an additional resource supporting visual and textual information. Visual video features provide the basic information for the video events or actions. Although it is difficult to obtain high levels of semantics by using visual information, a convincing way to construct an independent fully automated video annotation or action recognition model is to utilize visual information as the central resource. This way takes us to content-based video information retrieval.

Optical Flow Based Representation Content-based video information retrieval is the automatic annotation and retrieval of conceptual video items such as objects, actions, events, etc. using the visual content obtained from video frames. There are various methods to extract visual features and use them for different purposes. The visual feature sets they use vary from static image features (pixel values, color histograms, edge histograms, etc.) to temporal visual features (interest point flows, shape descriptors, motion descriptors, etc.). Temporal visual features combine the visual image features with the time information. Representing video information using temporal visual features generically means modeling the visual video information with temporal dimension. i.e., constructing temporal video information. We need to represent the temporal video information formally for developing video action recognition methods. Visual features such as corners, visual interest points etc. of video frames are the basics for constructing our model. These features are used for constructing a more complicated motion feature, namely, optical flow. In our work, we propose a new temporal video segment representation method to retrieve video actions for formalizing the video scenes as temporal information. The representation is fundamentally based on the optical flow vectors calculated for the frequently selected frames of the video scene. Weighted frame velocity concept is put forward for a whole video scene together with the set of optical flow vectors. The combined representation is used in the action based temporal video segment classification. Related Work

There are different approaches followed for the representation of temporal video segments for content-based video information retrieval problems such as video action recognition, event detection, cut detection, etc. The studies in [10, 11, 12, 13] focus on the perception of the visual world and bring us facts about how to detect the visual features and in which context more philosophically. Regarding the visual features, mentioned approaches can generally be figured out. Key-frame, bag-of-words, interest points and motion based approaches are the groups of approaches reflecting the way of representation. In our study, a motion-based representation is proposed to deal with the temporal video segmentation problem. Optical flow is the motion feature, integrating time with visual features, utilized for constituting the model. It is also important to mention the temporal video segmentation methods in the literature. Temporal video segmentation is the problem of temporally splitting the video into coherent scenes. It generally originated from the needs of video segment classification. In order to semantically classify the video scenes as segments, they need to be extracted considering all video information. Temporal video segmentation methods tackle the problem from different points of view. Because we are dealing with visual feature-based segmentation, we analyze and group the methods accordingly. The methods in question include pixel difference- based, histogram comparison-based, edge-oriented, and motion-based methods. Pixel difference-based methods use pixel intensities or color differences between the frames in order to characterize cuts between video scenes. A threshold-based automatic cut detection method is introduced in [12]. The method uses visual features for representing cut candidates and, according to these features, threshold values

are estimated. Despite its simplicity and time efficiency, the most important drawback of the pixel difference- based approach is its sensitivity to motion. The histogram comparison-based method uses color histogram differences between frames. It is especially successful in cases that are independent of motion. The most important drawback of this approach results from the meaning of the histogram itself. Histogram similarity, in many cases, does not mean real similarity in the context.

II. TEMPORAL SEGMENT REPRESENTATION

Temporal video segment representation is the problem of representing video scenes as temporal video segments. While this problem generally runs through the video information including visual, audio and textual features, our study deals with visual features only. Mentioned problem is originated from representing the temporal information. Temporal information provides a combined meaning composed of time and magnitude for a logical or physical entity. Robot sensor data, web logs, weather, video motion and network flows are common examples of temporal information. Independent from domain, both representation and processing methods of temporal information is important in the resulting models. Regarding the processing methods, prediction, classification and mining can be considered as first comers for the temporal information. In most cases, the representation is also a part of the processing methods due to the specific problem. While the representation and processing methods are handled together, the focus is especially on the processing methods rather than the representation in these cases. Temporal data mining and time series classification can be exemplified for the approaches on temporal information retrieval. The types of the features and their quality on describing the domain knowledge also influence the temporal information processing and its application. Also, having high dimensionality makes the effective representation of temporal information with more complicated features important. Therefore, feature definitions, construction and feature extraction methods play an important role in processing the temporal information. Optical Flow

Theoretically, optical flow is the motion of visual features such as points, objects, shapes etc. through a continuous view of the environment. It represents the motion of the environment relative to an observer. James Jerome Gibson firstly introduced the optical flow concept in 1940s, during World War II [16]. He was working on pilot selection, training, and testing. He intended to train the perception of pilots during the war. Perception was considered for the effect of the motion on the observer. In this context, shape of objects, movement of entities, etc. are handled for perception. During his study on aviation, he discovered optical flow patterns. He found that the environment observed by the pilot tends to move away from the landing point, while the landing point does not move according to the pilot. Region-Based Matching:

Region-based matching approaches alternate the differential techniques in case differentiation and numerical operations is not useful due to noise or small number of frames [17]. In region-based matching, the concepts such as velocity, similarity, etc. are defined between image regions. [21, 27] propose region-based matching methods for optical flow estimation. In [21], the matching is based on Laplacian pyramid while [27] recommends a method based on sum of squared distance computation. Energy-Based Methods:

Energy-based methods are based on the output energy of filters tuned by the velocity [17]. [26] proposes an energy- based method fitting spatiotemporal energy to a plane in frequency space. Gabor filtering is used in the energy calculations. Phase-Based Techniques:

Different from energy-based methods velocity is defined as filter outputs having phase behavior. [28, 23, 22] are the examples of phase-based techniques using spatiotemporal filters. Differential Techniques:

Differential techniques utilize a kind of velocity estimation from spatial and temporal derivatives of image intensity. They are based on the theoretical approach proposed by. The proposed approach results in the equation. Differential techniques are used for solving the problem generally represented by this equation. Horn- Shunck method is a fundamental method among the differential techniques. Global smoothness concept is also used in the approach. Lucas- Kanade method is also an essential method solving the mentioned differential equation for a set of neighboring pixels together by using a weighted window. Use second order derivatives generating the optical flow equations. Global smoothness concept is also used as well as the Horn-Shunck method. Proposes a distance based method efficient for real- time systems. The method is analyzed according to time-space complexity and its tradeoff. Suggests a classical differential approach. But, it is combined with correlation based motion descriptors. Optical Flow Based Segment:

Representation In this study, an optical flow based temporal video information representation is proposed. Optical flow vectors are needed to be calculated for the selected sequential frames. Optical flow estimation is important as the basic element of the model is optical flow vectors. In our approach, Shi-Tomasi algorithm proposed in is used for feature detection. As it is mentioned before, Shi-Tomasi algorithm is based on Harris corner detector and finds corners as interest points. Shi-Tomasi algorithm uses the eigenvalues of the Harris matrix. In this context, it differs from Harris corner detector. The algorithm assumes that minimum of two eigenvalues of Harris matrix determines the cornerness of the point. Therefore, the corner decision is done using the eigenvalues of the matrix. Shi-Tomasi algorithm gives more accurate results compared with Harris detector. The algorithm is also more stable for tracking. For estimating optical flow, Lucas-Kanade algorithm is selected. With videos having sufficient information and excluding noise, Lucas-Kanade algorithm is successful. The algorithm works for the corners obtained from Shi-Tomasi algorithm. Optical flow estimation: In our approach, the Shi–Tomasi algorithm proposed in is used for feature detection. The Shi–Tomasi algorithm is based on the Harris corner detector [27] and finds corners as interest points. The algorithm is especially robust for tracking. The Lucas–Kanade algorithm is selected [21] for estimation of the optical flow. The Lucas–Kanade algorithm is especially successful for videos with sufficient information and no noise. It works with the corners obtained from the Shi–Tomasi algorithm in our case. The following function should be minimized for each detected corner point, as seen in

differential approaches: E (δx,δy) = I (x, y) − I (x + δx, y + δy) (1) According to our optical flow implementation, video frames are selected according to a frequency of 6 frames/s for 30 fps videos. The implementation is first applied to the "Hollywood Human Actions" data set. Optical flow vectors are calculated for every detected point in all frequently selected frames. The set of optical flow vectors is the temporal information source for our representation. The model below forms the backbone of our representation method. The optical flow vector set with an operator constructs the representation. R = [S (V), Φ] S (V) is the set of optical flow vectors, while Φ is the descriptor operator. The operator defines the relation of the elements of the optical flow vector set of the frames. Conclusion:

In this study, we tried to solve a combination of different problems on action recognition. The fundamental problem inspires us is the representation of temporal information. In many fields, representation of temporal information is essential to retrieve information from a temporal data set. The solution to the problem varies from representing each temporal entity in a different time slice to representing a simple summary of the whole time interval. Efforts for finding a solution between these two endpoints, should try to tackle the problem from different point of views. This is because, the level of representation changes with the source of the problem. For instance, to represent all the information in all time slices for symbolizing the temporal information having high frequency over time, one should handle the curse of dimensionality problem. On the other hand, representing a single summary will cause the problem of lacking the flow of temporal information. In these cases, the focus of approaches will be finding supportive information from different sources and integration of these sources in a singular representation. We aimed to solve the temporal information representation problem in video domain. As the video information is a perfect example of high frequency temporal information, representation of video information is essential for the purposes based on video information retrieval. Video action recognition is selected as our specific domain. The problem domain is reduced to the temporal video segment classification. The study is shaped on visual features of the video information for the automaticity concerns. As it is mentioned below, the representation level determines the reduced problem. In this context, our aim is to represent the video scenes avoiding the lack of temporal information flow while without causing the curse of dimensionality problem. Therefore, using more descriptive and high level visual features having the ability to host the additional temporal nature of the simpler features such as color, edge, corner, etc. becomes unavoidable. This will pass the high load of temporal information residing in high dimensional representation to the mentioned high level features. An optical flow based approach is proposed in this paper for representing temporal video information by inspiring from the above studies. This generic approach is applied in both temporal video segment classification and temporal video segmentation. The adaptation of the model to video segment classification is presented. The weighted frame velocity concept is proposed to strengthen the representation with the velocity of video frames. This representation formalism is tested with SVM based classification of video segments. The results show that the proposed method produces encouraging results. The main advantage of the method is the multi-purpose temporal video representation model proposed for video action recognition domain. The new formalism described here is especially important for simplifying the computational complexity for high dimensional information.

ACKNOWLEDGMENT

I would like to thank Dr. Shalu Bashambu for her immense help and support, useful discussions and valuable recommendations.

REFERENCES

[1] T. C. Vasileios, C. L. Aristidis and P. G. Nikolaos, "Scene Detection in Videos Using Shot Clustering and Sequence Alignment"
[2] A. Ghoshal, P. Ircing and S. Khudanpur, "Hidden Markov Models for Automatic Annotation and Content Based Retrieval of Images and Video". Proc. of SIGIR, 2005.
[3] L. W. Chang, W. N. Lie, and R. Chiang, "Automatic Annotation and Retrieval for Videos". Proc. of PSIVT 2006, LNCS 4319, pp. 1030 – 1040, 2006.
[4] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video Event Classification Using String Kernels". Multimed Tools Application, 48:69–87, 2009.
[5] F. Wang, Y. Jiang and C. Ngo, "Video Event Detection Using Motion Relativity and Visual Relatedness". ACM Multimedia '08, October 26–31, 2008
[6] C. Ngo, T. Pong and H.Zhang, "Motion-Based Video Representation for Scene Change Detection". International Journal of Computer Vision 50(2), 127–142, 2002
[7] P. Sand and S. Teller, "Particle Video: Long-Range Motion Estimation Using Point Trajectories

