Image Guided Label Map Propagation in Video Sequences

Shuolin Di[1], Zhebin Zhang[1], Shiqi Wang[1], Nan Zhang[2], Siwei Ma[1]
[1]National Engineering Lab. For Video Tech., School of EECS, Peking University, Beijing 100871, China
[2]School of Biomedical Engineering, Capital Medical University
Email: {disl, zbzhang, sqwang, swma} @pku.edu.cn; zhangnan@ccmu.edu.cn

***Abstract*—In this paper, we propose a novel method to transmit the label maps by propagating from a key frame to non-key frames. The label map of a non-key frame is initialized by warping the label map of its corresponding key frame according to the motion estimation between them. Subsequently, the initialized label map is optimized with the guidance of its texture image. The optimization process minimizes an energy function which takes two constraints into consideration: (i) the data term measuring the similarity between an estimated label map and its initialized one, (ii) the regularization term enforcing the local smoothness in the label map and the consistency of region boundaries between the estimated label map and its corresponding texture image. Graph cuts based computation process is finally performed to generate the optimized label map. Experimental results show that our method achieves higher accuracy and better visual quality comparing with the state-of-the-art method.**

Keywords—label propagation; motion estimation; graph cuts

I. Introduction

Obtaining frame-based semantic label maps of a video sequence is one of the key steps in the typical computer vision and video processing tasks, e.g. object tracking, scene segmentation [1], geometry structure analysis [2], and even video compression [3]. Compared with the independently frame-by-frame label estimation, it not only reduces computational complexity but also maintains the temporal coherence by propagating label maps from key frames to the unlabeled non-key frames. Especially, when preparing training datasets in some learning tasks in video sequences, an automatic label propagation method could dramatically reduce the burden of manual annotation [4].

Intuitively, the label maps of non-key frames can be directly warped from their previous key frames according to the motion estimation between them [5]. However, such methods only work well with short-term reference but not for long-term one due to the accumulative estimating error and lack of correspondences [6]. For long-term reference, some probabilistic methods are proposed [4, 5-9]. In [6], a graphical model of HMM is proposed to couple image sequences and their annotation. The label maps of non-key frames are obtained by using the EM estimation method. J. Rituerto [7] introduces a superpixel based label propagating method, but the performance depends on the accuracy of superpixel extraction. Other learning based methods are also proposed, e.g. by using localized random forests [8] or temporal tree structure [9]. Bai and Sapiro [10] treat the video as a space-time volume and propagate labels via the shortest geodesic distance. All the aforementioned methods heavily depend on the complex appearance models while avoiding the use of dense optical flow. In [4], a probabilistic pixel labeling model is proposed by combing motion, appearance and spatial smoothness constraints. Comparing with the methods which use motion or appearance information alone, this method achieves better performance. However, the performance of this method depends on the assumed distributions of label uncertainties. Thus, it is hard to determine the adaptive probabilistic distribution for different videos. Another disadvantage of the learning based methods is that the applications of such methods are constrained by the training data.

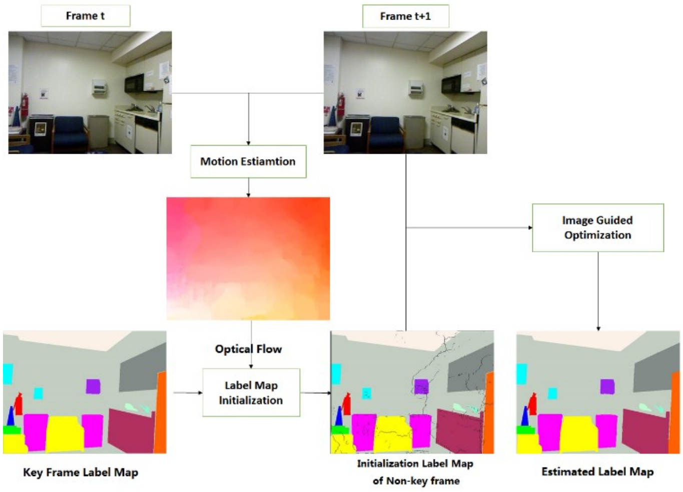


Fig. 1. The framework of the proposed method.

To address the aforementioned issues, we propose a label map propagation method which combines the constraints both from motion estimation and texture information. The label map of a key frame can be provided either by manual annotations or by the estimation using certain trained classifier/estimator. The label map of a non-key frame is estimated by solving an optimization problem on the defined energy function by graph cuts. The proposed energy function consists of two terms, (i) the data term to evaluate the label map similarity between the key frame and the estimated one and (ii) the regularization term to maintain the spatial structure coherence between the estimated label map and its corresponding texture image.

978-1-4799-8391-9/15/$31.00 ©2015 IEEE

As shown in Fig. 1, the data term is computed by comparing the estimated label map with the initialized one, and the initialized one is directly warped from the key frame's label map according to the motion information. The regularization term is calculated by introducing an anisotropic total generalized variation, which provides the underlying description of the region boundaries in the texture image and its local smoothness inner the regions.

Instead of using complex appearance model, the proposed method utilizes the constraint of structure consistency between the estimated label map and its corresponding texture image. Thus it is adaptive to more scenarios in real applications without the constraint of training dataset in the learning based methods. Since the coherence between the estimated label map and texture image is maintained, our method overcomes the shortcomings of motion estimation based propagation.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed label map propagation method. The experimental results are presented and discussed in Section 3. Section 4 concludes this paper.

II. IMAGE GUIDED LABEL MAP PROPAGATION

Given a video shot, the label map of a key frame (the first frame in this paper) is annotated by handcraft. We propose a label map estimation model which utilizes the corresponding texture information to propagate the label map from the key frame to non-key frames. The flowchart of the proposed method is illustrated in Fig. 1. This method estimates the label map of the current frame by utilizing both motion estimation for initialization and texture image guidance in the optimization. The label maps of the video sequence can be obtained frame by frame with the proposed method.

We formulate the label map propagation as a discrete optimizing problem. The optimized estimated label map L_E is obtained by minimizing an energy function as follows,

$$L_E = \underset{u}{\arg\min} \left\{ D(u, L_S) + \alpha S(u) \right\} \tag{1}$$

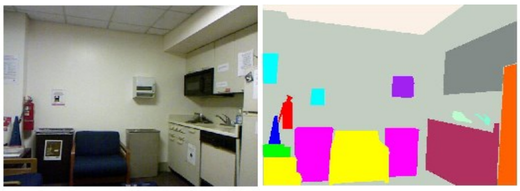
where the energy function consists of two term, the data term D and the regularization term S. The data term measures the similarity of an estimated label map u and the initialized one, L_S. The regularization term reflects the priors of the label map's smoothness among the neighboring pixels and similarity of region boundary between the label map and its texture image. The factor α is used to balance the relative weight between two terms.

A. Data Term

The data term ensures the consistency between the estimated label map and the initialized one, which is formulated as follows,

$$D(u, L_s) = \int_\Omega W(x) \cdot \delta\big(u(x), L_s(x)\big) dx \tag{2}$$

where Ω is the image space of non-key frame, x denotes a pixel coordinate and W denotes a weighting matrix. L_S is the initial label map warped from the key frame's label map according to the motion estimation between them. The function δ is used to penalize the difference between the label of pixel x and that of pixel y and it is defined as



(a) Texture image (b) Manual labeled map

Fig. 2. Texture and manual-labeled map.

$$\delta(x, y) = \begin{cases} 1, x \neq y \\ 0, x = y \end{cases} \tag{3}$$

In our implementation, we utilize the optical flow method mentioned in [11] to obtain the motion estimation, which employs the state-of-the-art algorithms and optimizing method. The weighting matrix w in Eqn (2) is equal either to zero or one. If there is an initial label value at pixel x which is warped from the key frame, w is set to one. Otherwise, w is set to zero.

B. Regularization Term

In order to optimize the initial estimation, the regularization term is added to the energy function to constrain the structure similarity between the estimated label map and the texture image, i.e. region boundary coherence and inner-region smoothness. As shown in Fig. 2, pixels in the smooth region are correlative and should be labeled as the same one. Pixels among the texture boundary have lower correlation and should be labeled independently.

In this paper, the regularization term is formulated as

$$S(u) = \sum_{x,y \in N} B_{\{x,y\}} \cdot \delta\big(u(x), u(y)\big) \tag{4}$$

where N is the set of all pairs of neighboring pixels. The coefficient $B_{\{x,y\}}$ describes the relative importance of the interaction between neighboring pixels x and y.

Assuming that texture edge most likely correspond to estimating label discontinuities, the coefficient $B_{\{x,y\}}$ can be interpreted as a descriptor for a discontinuity between x and y. Normally, $B_{\{x,y\}}$ is large when pixel x and y are similar (e.g. in their intensity) and $B_{\{x,y\}}$ is close to zero when they are different. In this way, the difference in labeling between pixels in smooth region will be assigned as a penalty. In addition, pixels among the texture boundary can be labeled independently.

We introduce an anisotropic diffusion tensor to capture the edge information in the image. By expressing the coefficient $B_{\{x,y\}}$ as the anisotropic diffusion tensor, we can penalize estimating discontinuities at homogeneous regions and allow sharp edges at corresponding texture differences. Further, the regions where the label is interpolated are filled out reasonably. The tensor is formulated as

$$B = \exp\left(-\beta \left|\nabla I_H\right|^\gamma\right) nn^T + n^\perp n^{\perp T} \tag{5}$$

where n is the normalized direction of the image gradient $n=\nabla I_H/|\nabla I_H|$, $n^{\perp}$ is the normal vector to the gradient and the factors β,γ adjust the magnitude and the sharpness of the tensor.

C. Optimization

According to the definition of the data term and the regularization term, the energy function is then reformulated as follows,

$$\min_{u}\left\{\alpha\sum_{x,y\in N}B\cdot\delta\big(u(x),u(y)\big)+\sum_{\Omega}W(x)\cdot\delta\big(u(x),L_s(x)\big)\right\} \quad (6)$$

which consists of the data term and the regularization term with anisotropic diffusion. Image labeling could be modeled as Markov Random Field. The optimization of the proposed energy function is a discrete energy minimization on MRF. Graph cuts technique can be used to solve this problem and usually very fast [12]. The main idea of graph cuts is to define a graph such that there is a one-to-one correspondence between configurations of the energy function and cuts of the graph. And the total cost of the cut is exactly the same as the total energy of the proposed function. As illustrated in [13], the proposed energy function can be minimized by graph cuts. The data and regularization term in the energy function can be mapped to data cost and smooth cost in the graph cuts correspondently. The details of the construction of multi-label graph and the optimizing process can be found in [14].

As mentioned above, error accumulation introduces some estimating error. Motivated by the approach used in video coding standard, we divide the whole video sequence into separate groups of frames. In this paper, each group has 5 frames and the non-key frame refers to the first frame of the group to calculate optical flow estimation.

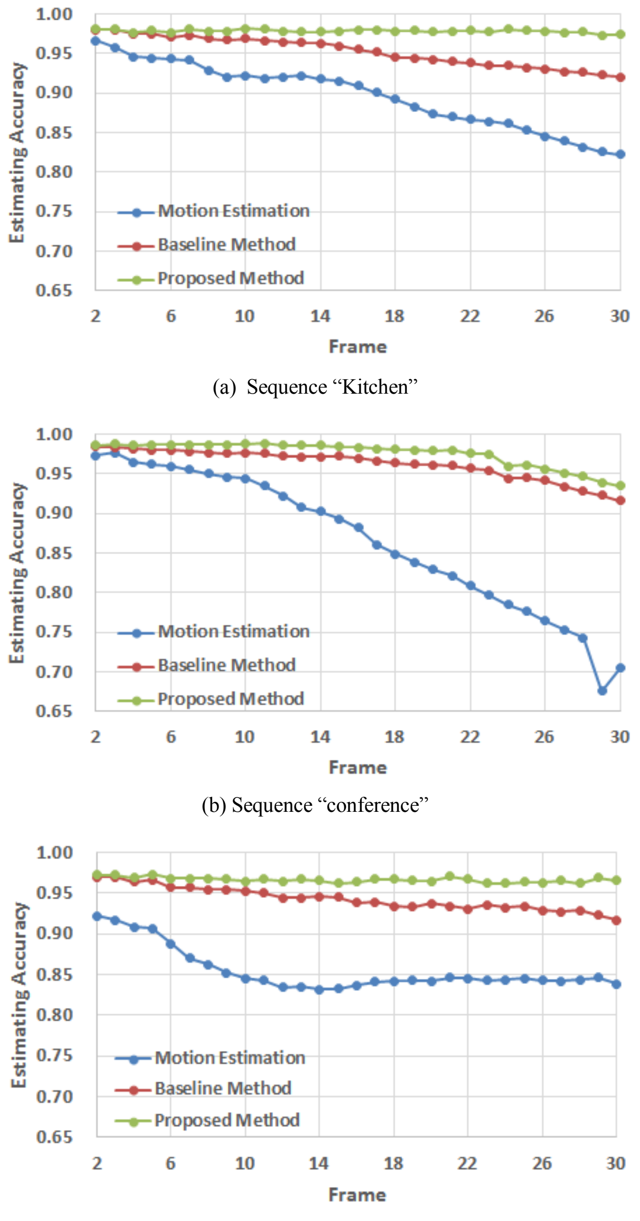
III. Experimental Results

A. Dataset and Implementation

We evaluate the proposed method on the NYU dataset [15]. This dataset comprises of video sequences from a variety of indoor scenes and contains abundant and complex indoor objects. The sequences are recorded by Microsoft Kinect that consists of both the texture and depth sensors. However, merely frames are provided with label map in the NYU dataset. In order to evaluate our method, we utilize MIT LabelMe online annotation tool [16] to further label the frame in three video sequences. Referring to the label maps given by NYU dataset, the ground truth is annotated by volunteers using LabelMe.

B. Experimental Result and Analysis

Fig. 4 shows several estimated label maps sampled from 30 propagated label maps of sequences “Kitchen” and “Conference”. We compare the proposed method with the motion estimation method as well as the baseline method [4]. Motion based method warps the label map of key frame by directly using motion estimation and may produce holes. These holes form because the correspondence established by motion estimation between frames is neither injective nor surjective. Guided by the image clues, the proposed method accurately fills the hole. Comparing to the baseline method [4], the proposed method maintains the consistency of the region boundaries and inner region smoothness between the estimated label map and its corresponding texture image very well. Because the regularization term in the energy function is used to maintain the spatial structure coherence. Referring to the ground truth, performance comparison of different methods is made. As plotted in Fig. 3, comparing to the baseline method, the proposed method achieves higher estimation accuracy with better visual quality and maintains a stable performance over the test video sequences. As shown in Fig. 4 (b), the proposed method lacks the ability of recognizing new appearing objects. In Fig. 3 (b), the performance is degraded correspondingly.

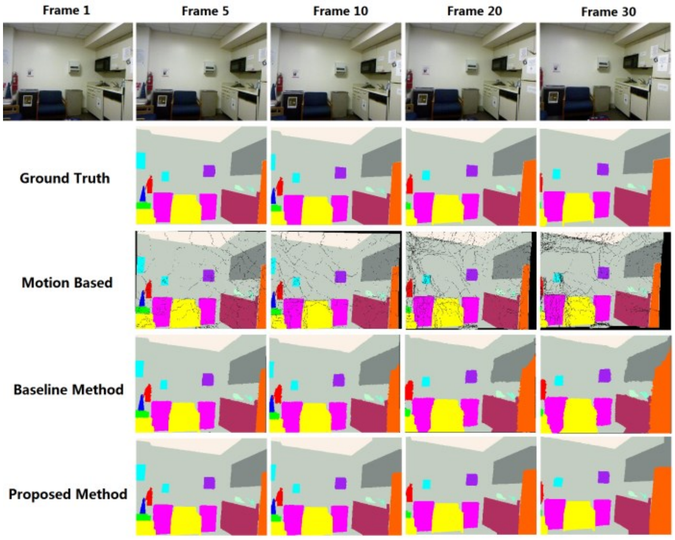


(a) Sequence “Kitchen”

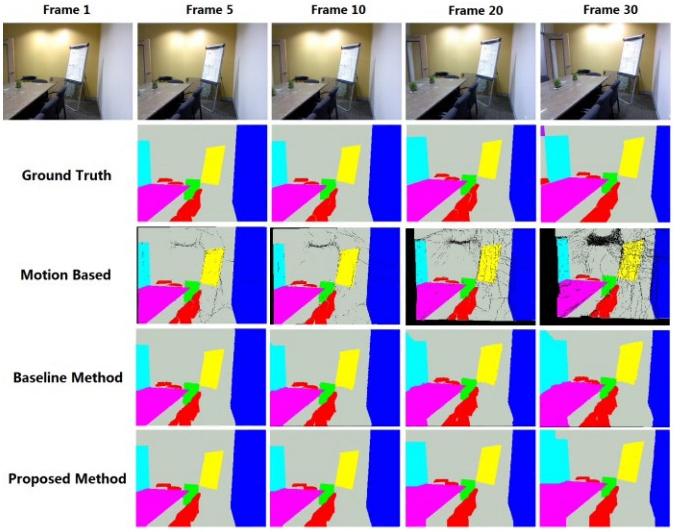
(b) Sequence “conference”

(c) Sequence “bookstore”

Fig. 3. Performance comparison of label propagation in terms of different method.



(a) Sequence "Kitchen"



(b) Sequence "Conference"

Fig. 4. Qualitative comparison of label propagation by different methods.

IV. CONCLUSION

In this paper, we propose a label map propagating method with image guidance. The proposed model considers both the spatial constraints and temporal information. The experimental results demonstrate that our method achieves higher estimation accuracy and better performance compared with the baseline method.

It is noted that, if the label map is propagated across different video shots or there are new objects appearing in the non-key frame, the performance would be degraded. In the future work, we will investigate this problem and propose robust solutions for frame label propagation.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China (61300062, 61210005, 61322106, 61421062, 61272255 and 61231010), the Major State Basic Research Development Program (2015CB351800) and the International Postdoctoral Exchange Fellowship Program of the Office of China Postdoctoral Council under Grant 20130008. This work was also granted by Cooperative Medianet Innovation Center.

REFERENCES

[1] Z. Zhang, C. Zhou, W. Gao, and W. Yizhou, "Interactive stereoscopic video conversion," CSVT, IEEE T. on, Vol.23(10), pp.1795-1808, 2013.

[2] J Zhang, Z Zhang, "Depth Map Propagation with the Texture Image Guidance" IEEE Proceeding of the 21st International Conference on Image Processing (ICIP2014), pp. 3813-3817, Paris, October, 2014.

[3] Z. Zhang, C. Zhou, R. Wang, Y. Wang, and W. Gao, "A compactrepresentation for compressing converted stereo videos," Image Processing, IEEE T. on, Vol.23(5), pp.2343-2355, April, 2014.

[4] A. Y. C. Chen, J. J. Corso, "Propagating multi-class pixel labels throughout video frames," Image Processing Workshop (WNYIPW), 2010 Western New York, IEEE, pp.14-17, 2010.

[5] A. Fathi , M. F. Balcan, X. Ren, "Combining self training and active learning for video segmentation," Proceedings of the British Machine Vision Conference, 2011, Vol. 29, pp. 78.1-78.11,2011

[6] V. Badrinarayanan, F. Galasso, R. Cipolla, "Label propagation in video sequences," 2010 IEEE Conference on. IEEE, Computer Vision and Pattern Recognition (CVPR), pp. 3265-3272, 2010.

[7] J. Rituerto, A. C. Murillo, Jana Kosecka, "Label propagation in videos indoors with an incremental non-parametric model update," Intelligent Robots and Systems (IROS), 2011.

[8] Naveen Shankar Nagaraja, Peter Ochs, Kun Liu, Thomas Brox, "Hierarchy of localized random forests for video annotation," Pattern Recognition, pp. 21-30, 2012.

[9] V. Badrinarayanan, I. Budvytis, R. Cipolla, "Semi-supervised video segmentation using tree structured graphical models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, , 35(11), pp. 2751-2764, 2013.

[10] X. Bai, G. Sapiro, "Geodesic matting: A framework for fast inter-active image and video segmentation and matting," IJCV, 2009.

[11] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," Doctoral Thesis. Massachusetts Institute of Technology May. 2009.

[12] J. H .Kappes, B. Andres, F. A. Hamprecht, "A comparative study of modern inference techniques for discrete energy minimization problems," Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, pp. 1328-1335， 2013.

[13] Y. Boykov, V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 26(9): 1124-1137, 2004.

[14] A. Delong, A. Osokin, H. N. Isack, "Fast approximate energy minimization with label costs," International journal of computer vision, 96(1): 1-27, 2012.

[15] N. Silberman, D. Hoiem, P. Kohli, and R Fergus, "Indoor segmentation and support inference from RGBD images," Computer Vision–ECCV 2012, Springer Berlin Heidelberg, pp. 746-760, 2012.

[16] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," International Journal of Computer Vision, Vol. 77, pp. 157-173, Numbers 1-3, May,2008.