MULTI-LAYER TEMPORAL GRAPHICAL MODEL FOR HEAD POSE ESTIMATION IN REAL-WORLD VIDEOS

Meltem Demirkus, Doina Precup, James J. Clark, Tal Arbel

Centre for Intelligent Machines,
Department of Electrical and Computer Engineering, McGill University

ABSTRACT

Head pose estimation has been receiving a lot of attention due to its wide range of possible applications. However, most approaches in the literature have focused on head pose estimation in controlled environments. Head pose estimation has recently begun to be applied to real-world environments. However, the focus has been on estimation from single images or video frames. Furthermore, most approaches frame the problem as classification into a set of coarse pose bins, rather than performing continuous pose estimation. The proposed multi-layer probabilistic temporal graphical model robustly estimates continuous head pose angle while leveraging the strengths of multiple features into account. Experiments performed on a large, real-world video database show that our approach not only significantly outperforms alternative head pose approaches, but also provides a pose probability assigned at each video frame, which permits robust temporal, probabilistic fusion of pose information over the entire video sequence.

***Index Terms*—** Head pose, real-world video, local invariant feature, probabilistic, graphical model.

1. INTRODUCTION

Head pose has been used as prior or contextual information in many applications, such as human computer interaction, face recognition, face verification and facial attribute classification. Considering the recent interest in real-world unconstrained videos (e.g. surveillance data) and the wide range of possible applications, robust and automatic head pose estimation from 2D images has been receiving such much attention [2, 3, 4, 5, 6, 7, 8, 9]. For example, in [4] head pose is used to map real-world face images to a common coordinate system.

The literature on head pose estimation from 2D images can grouped as: appearance template, manifold, subspace embedding, geometric and tracking based methods [10]. However, many of these approaches rely on requirements or assumptions that are not feasible in the context of unconstrained environments: (i) assuming that the entire set of facial features typical for frontal poses is always visible, (ii) manually labeling facial features in the testing data, (iii) relying on a known initial head pose for video sequences, which must be reinitialized (at times, manually) whenever the tracking fails, (either due to a failure in the face detection or due to occlusion) [10]. Some of the recent approaches developed for real-world environments treat head pose estimation as a classification task [5, 7]. That is, assigning a face image to one of very coarsely defined discrete poses (pose bins). Furthermore, they use single and low resolution video frames collected from crowded scenes under poor lighting, although some use relatively higher quality video frames/images and perform classification on finer pose bins [3, 6]. Some other works, on the other hand, define the pose estimation problem as a continuous pose angle estimation task [8, 9]. Most of these approaches either focus on only one set of features to represent faces, or they do not leverage the temporal pose information available in video sequences in their frameworks. Our hypothesis is that by using complementary, robust local invariant features, and leveraging the dependencies between the frames in the video sequence, one can substantially improve head pose estimation in real world scenarios.

This paper addresses the problem of automatic continuous

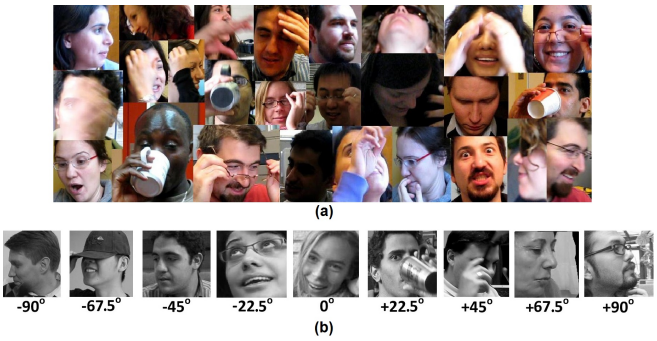

Fig. 1. Sample face images from McGill Real-World Face Video Database: (a) Challenges of real-world environment include wide variability in illumination conditions and background clutter, arbitrary head poses and scales, arbitrary partial occlusions etc., (b) Head pose (yaw angle) labels of sample face images provided by the probabilistic labeling strategy in [1].

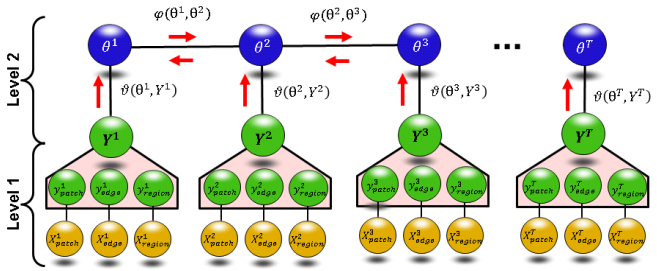

Fig. 2. An overview of the proposed graphical model.

head pose (yaw angle) estimation in real-world videos which consider the joint occurrence of arbitrary face scales, extreme head poses, non-uniform illumination conditions, partial occlusions, motion blur, background clutter, wide variability in image quality, and subject variability (Figure 1(a)). The proposed multi-layer and temporal graphical model (Figure 2) uses spatial codebook representations obtained from different local invariant features which have a high degree of invariance to various transforms, such as changes in scale, viewpoint, rotation and translation. The features are chosen in a fashion such that they extract complementary information from the tracked face image: (i) facial edge points obtained from eyebrow, mouth etc., (ii) facial anatomical regions extracted from eyes, forehead, cheeks etc., and (iii) densely sampled patch-based features over the whole face image (see Figure 3(a)). The codebook statistics are used to calculate the pose distribution for each feature type using Random Forests (RF) [11], which are later used in a graphical model to estimate the single video frame pose probability distribution. Next, the framework temporally models these continuous head pose probabilities over a video sequence using Belief Propagation (BP) [12]. Finally, the method performs density estimation over the discrete head pose probabilities inferred. The experiments are performed on a large, challenging, public video database, namely the McGill Real-World Face Video Database [1] (see Figure 1) . The probabilistic labeling module in [1] is employed to collect the pose ground truth (leading to 9 MAP pose estimates (bins) (see Figure 1(b))), which are used for the classification tasks . The results show that that the proposed framework outperforms alternative approaches [3, 9, 13, 8, 6].

2. METHODOLOGY

The proposed framework has two main levels (see Fig. 2). As a preprocessing step, the robust algorithm in [1] is used to locate and track faces in real-world video sequences. Then in Level 1, complementary local invariant facial features are extracted from the detected faces. The relationship between features extracted and their corresponding pose beliefs are modelled. In Level 2, for each single frame, the head pose distribution is inferred based on the different feature-based pose beliefs inferred at Level 1. The temporal information is leveraged to estimate the most likely head pose configuration, which can be achieved through BP.

2.1. Level 1: Estimation of Pose Distribution from Single Video Frames

We assume that a video sequence (or clip) contains T video frames. Once the pre-processing steps are complete (e.g. tracking, detection), the face is assumed to be properly detected and localized in each frame. Each tracked face can be represented using a variety of complementary representations (e.g. patches, edges, regions), and then modelled using a variety of local invariant feature detectors/descriptors. One can use different types of local invariant feature detectors/descriptors to model each face representation. However, it is crucial to choose complementary representations which could provide high accuracy in terms of pose classification performance when all features are combined (see Section 3 for selected features). Facial image patches, for example, achieve dense sampling and modelling of the facial characteristics whereas the facial edges model the facial lines, such as the eyebrow line and mouth line. Facial regions, on the other hand, model the anatomical regions, such as eye, eyebrow, mouth, nose and cheek. Once these features are detected and descriptors are extracted, corresponding spatial-Bag-of-Words (BOW) representations are learned. Rather than using the patch index as the spatial information as done in [1], here we directly use the extracted feature location in the coding and pooling phases. Because faces are aligned in the preprocessing step, this mapping provides better modelling for the face vocabulary. Finally, each face image at frame t is represented by the codebook occurrences for each corresponding feature type, i.e. X^t_{patch}, X^t_{region} and X^t_{edge}. Once each frame is represented by patch, edge and region based codebook occurrence statistics, for each codebook type, a RF [11] is trained to estimate the corresponding pose distributions: $\left\{y^t_{patch}, y^t_{edge}, y^t_{region}\right\}$. The motivation behind the use of RFs is due to its high generalization power, fast computation, ease of implementation, embedded feature selection property, and its high classification performance [14].

2.2. Level 2: Estimation of Video Pose Distribution

Given the patch, edge and region based pose probability distributions for a video frame at time t, i.e. $Y_t = \left\{y^t_{patch}, y^t_{edge}, y^t_{region}\right\}$, we want to estimate the probabilities for a set of head pose angles $\theta_t = \{\phi_1, \phi_2, \cdots, \phi_M\}$. The ultimate goal is to estimate an entire set of pose probability density functions throughout a video, i.e. $\Theta = \{\theta_1, \theta_2, \cdots, \theta_t, \cdots, \theta_T\}$. The posterior distribution is $p(\Theta|\vec{Y}) = \frac{p(\Theta,\vec{Y})}{p(\vec{Y})}$, where $\vec{Y} = \{Y_1, Y_2, \cdots, Y_t, \cdots, Y_T\}$ and $p(\vec{Y})$

is a normalization constant Z with respect to Θ, such that $p(\Theta|\vec{Y}) = \frac{1}{p(Z)}p(\Theta, \vec{Y})$. Note that, if Z can not be calculated directly, $p(\Theta, \vec{Y})$ becomes an approximation. Computing this posterior distribution can be difficult without any approximations [15]. Thus, we use the graphical model shown in Figure 2 to model the head pose over a video sequence Θ. So, the posterior distribution is expressed as an MRF with pairwise interactions: $p(\Theta|\vec{Y}) = \frac{1}{Z}\left(\prod_{t=1}^{T}\vartheta(\theta^t, Y^t)\right).$ $\left(\prod_{t=1}^{T-1}\varphi(\theta^t, \theta^{t+1})\right)$, where $\vartheta(\theta^t, Y^t)$ is the unary compatibility function accounting for local evidence (likelihood) for θ^t and $\varphi(\theta^t, \theta^{t+1})$ is the pairwise compatibility function between θ^t and θ^{t+1}.

One way to estimate the most likely head pose configuration is by calculating the MAP estimate, i.e. $\Theta^* = \text{argmax}_\Theta p(\Theta|\vec{Y})$, which can be achieved through BP [12]. BP is an inference method developed for graphical models, which can be used to estimate the *marginals* or the most likely *states*. In our experiments, we adapt the "sum-product" BP algorithm which estimates the probability distributions. BP provides the exact solution if there is no loop (cycle) in the graph, i.e. if the graph is a chain or a tree [12], which is the case here. In order to estimate the marginal distributions, the BP algorithm creates a set of message variables which are updated iteratively via passing between neighbours. $m_{t,t+1}(\theta_{t+1})$ corresponds to the message sent from node t to node $t+1$ about the degree of its belief that node $t+1$ should be in state θ_{t+1} (see Figure 2). The BP algorithm updates the messages according to:

$$m_{t,t+1}^{(q+1)}(\theta^{t+1}) = \frac{1}{Z_{t+1}}\sum_{\theta^t}\varphi(\theta^t, \theta^{t+1})\vartheta(\theta^t, Y^t) \prod_{k\in N(t)\setminus t+1} m_{k,t}^{(q)}(\theta_t) \quad (1)$$

where $\frac{1}{Z_{t+1}} = \sum_{\theta^t} m_{t,t+1}^{(q+1)}(\theta^t)$ is a normalization factor, and the set of nodes in the neighbourhood of t is denoted by $N(t)$. $(q+1)$ and (q) represent the iteration indices. The initial messages $m_{t,t+1}^{(0)}(.)$ are typically initialized to uniform positive values. In a general graph, the update procedure is repeated iteratively until the messages converge to a consensus, then the *marginals (beliefs)* are calculated (Equation 2). Since our graph here is acyclic, two passes are sufficient to compute all messages, making the algorithm efficient. The *belief* (b_t) is an estimate of the marginal distribution, derived from converged message variables as follows:

$$b_t(\theta^t) = \frac{1}{\tilde{Z}_t}\vartheta(\theta^t, Y^t)\prod_{k\in N(t)} m_{k,t}(\theta^t) \quad (2)$$

where $\tilde{Z}_t$ is a normalization factor guaranteeing that $\sum_{\theta^t} b_t(\theta^t) = 1$. Since our graph does not have loops, the *beliefs* are guaranteed to be the true marginals $p(\theta^t|\vec{Y})$. Note that in the case of "sum-product" BP, the belief is an estimate of marginals whose maximal point indicates the most likely state. The pairwise compatibility function $\varphi(\theta^t, \theta^{t+1})$ is assumed to be a Gaussian distribution $N(\mu, \Delta)$ with mean μ and covariance matrix Δ. Furthermore, we define the unary compatibility function for each node *i*, i.e. $\vartheta(\theta^t, Y^t)$, as the joint distribution $p(\theta^t, Y^t) = p(\theta^t, y_{patch}^t, y_{edge}^t, y_{region}^t)$, which is equal to:

$$p(\theta^t, y_{patch}^t, y_{edge}^t, y_{region}^t) = p(\theta^t|y_{patch}^t, y_{edge}^t, y_{region}^t)\cdot p(y_{patch}^t, y_{edge}^t, y_{region}^t) \quad (3)$$

where the posterior probability is $p(\theta^t|y_{patch}^t, y_{edge}^t, y_{region}^t) = \frac{1}{\mathcal{Z}(y_{patch}^t, y_{edge}^t, y_{region}^t)}exp\{-U\}$, where as before $\mathcal{Z}$ denotes the normalization function and the energy function U is defined as:

$$\begin{aligned} U = {} & \beta_1\nu(\theta^t, y_{patch}^t) + \beta_2\nu(\theta^t, y_{edge}^t) + \\ & \beta_3\nu(\theta^t, y_{region}^t) + \beta_4\nu(\theta^t, y_{patch}^t, y_{edge}^t) + \\ & \beta_5\nu(\theta^t, y_{patch}^t, y_{region}^t) + \beta_6\nu(\theta^t, y_{edge}^t, y_{region}^t) + \\ & \beta_7\nu(\theta^t, y_{patch}^t, y_{edge}^t, y_{region}^t) \end{aligned} \quad (4)$$

where $\{\beta_1, \cdots, \beta_7\}$ are weights, which are learned on the training data using 2-fold cross validation. The potential function ν models the possible cliques of t-th frame pose distribution with estimates from three different feature representation, such as *pairwise* (e.g., (θ^t, y_{patch}^t)) and *triplet* (e.g., $(\theta^t, y_{patch}^t, y_{edge}^t)$) cliques. ν is defined by the corresponding probability distribution functions:

$$\nu(\theta^t, y_{patch}^t) = -\log\left\{p(\theta^t|y_{patch}^t)p(y_{patch}^t)\right\} \quad (5)$$

$$\nu(\theta^t, y_{patch}^t, y_{edge}^t) = -\log\left\{p(\theta^t|y_{patch}^t, y_{edge}^t)p(y_{patch}^t, y_{edge}^t)\right\} \quad (6)$$

$$\nu(\theta^t, y_{patch}^t, y_{edge}^t, y_{region}^t) = -\log\left\{p(\theta^t|y_{patch}^t, y_{edge}^t, y_{region}^t)p(y_{patch}^t, y_{edge}^t, y_{region}^t)\right\} \quad (7)$$

The probabilities $p(y_{patch}^t)$, $p(y_{patch}^t, y_{edge}^t)$ and $p(y_{patch}^t, y_{edge}^t, y_{region}^t)$ can either be assumed to be uniform or they can be calculated using the training database. For instance, $p(y_{patch}^t, y_{edge}^t) \propto k(y_{patch}^t, y_{edge}^t) + d_t$, where $k(y_{patch}^t, y_{edge}^t)$ is the count of the joint occurrence event $(y_{patch}^t, y_{edge}^t)$, and d_t is the Dirichlet regularization parameter required to compensate for the sparsity. Because a uniform prior is assumed, d_t is constant for all t. Note that other *pairwise*, *triplet* and *quadruplet* probabilities, for all combinations, is calculated in a similar fashion. The posterior probabilities $p(\theta^t|y_{patch}^t)$, $p(\theta^t|y_{patch}^t, y_{edge}^t)$ and$p(\theta^t|y_{patch}^t, y_{edge}^t, y_{region}^t)$, on the other hand, are calculated using the RFs [11]. Next, the entire pose density is estimated for 1^o intervals in the range $[-90^o, +90^o]$. Gaussian kernel-based model fitting is used since the initial pose densities do not follow any specific parametric distribution.

3. IMPLEMENTATION AND EXPERIMENTS

In our experiments, we use the challenging McGill Real-World Face Video Database [1] consists of 18,000 video

Table 1. Comparison of the different pose classification approaches over all folds (mean $\pm$ std).

	Accuracy (%)
Aghajanian and Prince [9] (BMVC'09)	20.68 ± 3.55
BenAbdelkader [13] (ECCV'10)	15.50 ± 2.73
Demirkus et al. [6] (ICIP'11)	40.12 ± 7.45
Demirkus et al. [8] (CVPRW'12)	55.04 ± 6.53
Zhu and Ramanan [3] (CVPR'12)	57.49 ± 0.91
Proposed Model	**73.09±3.61**

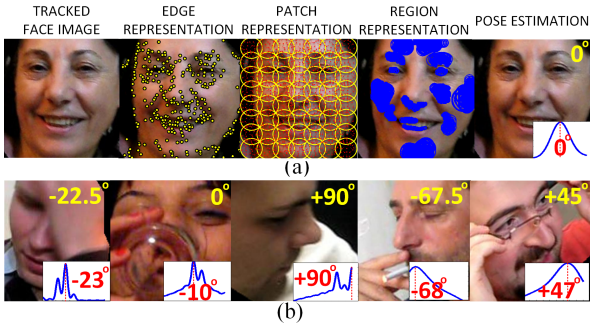

Fig. 3. Sample tracked faces images depicting the extracted facial features and pose estimates: (a) Location of some of the sample facial edge (GB), patch (SIFT) and facial region (BPLR) features. On the top right, the pose ground truth label (in yellow) is obtained via [1]. On the top bottom, the pose distribution (in the range of $[-90^o, +90^o]$) calculated by the proposed approach is in blue, and the MAP estimate is in red, (b) Example tracked face images with the corresponding pose ground truth labels, estimated pose distributions and the MAP estimates.

frames from 60 unconstrained videos of different subjects. Individual video frames exhibit wide variability in head pose: 45.8% of the frames are beyond 0^o, i.e. non-frontal, of which 58.1% go beyond $\pm 45^o$. Although in this paper the head pose density function is estimated for each pose angle at each frame in the sequence, we are bounded by the precision of the manual labeling of the database. Since it is not possible to collect the "absolute head pose angle" label from fully uncontrolled real-world data, we use the robust 2-stage labeling strategy introduced in [1] to collect ground truth pose labels. This leads to 9 pose labels (see Figure 1(b)) which are used in the evaluation phase. Furthermore, the faces are tracked and localized through the tracking algorithm explained in [1].

SIFT [16], Geometric Blur (GB) [17] and Boundary-preserving Local Region (BPLR) [18] features are used in our graphical model due to their high performance. The patch representation achieves a dense sampling and models each image patch with robust SIFT features [16]. To detect the key facial edge points and to calculate the corresponding descriptor around each edge point, the GB framework in [17] is used. Anatomical regions, such as the mouth, the eyes, the ears and the eyebrows, can provide some pose information. Local region detectors and descriptors are used to model anatomical regions. To achieve this, boundary-preserving local regions (BPLRs) [18] are chosen. BPLRs are densely sampled local regions obtained from a given face image, and they preserve the shape of the facial structure on which they are detected (for details, see [16, 17, 18]).

To compare the performance of the proposed approach against alternative approaches, we perform 10-fold cross validation on the McGill Real-World Face Video Database. For the patch-based probabilistic regression framework in [9] and the Bayesian models in [8, 6], the MAP of the probability density function serves as the estimated angle. The implementation and the proper training parameters for [3, 9] and [8, 6] are provided by the authors. Thus, the reported results are not affected by errors in implementation or in algorithm learning step. Since [9, 13, 8] provide continuous pose estimation, their pose spaces are discretized to be able to compare their accuracy with the other approach. In Table 1, the mean and standard deviation (std) statistics over the head pose classification performance for different approaches over all folds are provided. BenAbdelkader's supervised manifold-based approach [13] is chosen since manifold learning methods are reported to provide the highest head pose accuracy [10]. However, it performs the poorest since pose manifold is created using pixels intensities, which are not the optimal features for the real-world environments. The work by Zhu and Ramanan [3], which is a unified model for face detection, pose estimation and landmark localization using a mixture of trees with a shared pool of parts, provides the best accuracy (57.49%) among the competitors. The proposed framework, on the other hand, "significantly" (p-value of 2.8137e-13) outperforms the comparable approaches. The high classification accuracy (73.09%) is achieved using the probabilistic temporal model which takes the advantage of multiple features to model head pose distribution. Figure 3 shows some qualitative results along with estimated pose distributions, MAP estimates and the pose ground truth labeling obtained using [1]. It is observed that in the presence of bad face tracking and motion blur, the proposed model might fail due to the lack of reliable facial features.

4. CONCLUSIONS AND FUTURE WORK

We proposed a novel multi-layer temporal graphical model to use multiple features to infer continuous head pose angle from real-world videos, unlike most approaches. Experiments performed on a large real-world video database showed that our approach significantly outperformed the state-of-the-art approaches. We are currently investigating how to incorporate feature-level fusion with the current classifier-level fusion model to further improve our pose estimation performance.

5. REFERENCES

[1] M. Demirkus, J. J. Clark, and T. Arbel, "Robust semi-automatic head pose labeling for real-world face video sequences," *Multimedia Tools and Applications*, pp. 1–29, 2013.

[2] G. Hua, M.H. Yang, E.G. Learned-Miller, Y. Ma, M.A. Turk, D.J. Kriegman, and T.S. Huang, "Special section on real-world face recognition," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 33, no. 10, 2011.

[3] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 2879–2886.

[4] N. Kumar, A. Berg, P.N. Belhumeur, and S.; Nayar, "Describable visual attributes for face verification and image search.," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 33, pp. 1962–1977, 2011.

[5] J. Orozco, S.G. Gong, and T. Xiang, "Head pose classification in crowded scenes," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2009.

[6] M. Demirkus, B. N. Oreshkin, J. J. Clark, and T. Arbel, "Spatial and probabilistic codebook template based head pose estimation from unconstrained environments," in *ICIP*, Benoît Macq and Peter Schelkens, Eds. 2011, pp. 573–576, IEEE.

[7] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani, "Multi-class classification on riemannian manifolds for video surveillance," in *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, 2010, pp. 378–391.

[8] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, "Soft biometric trait classification from real-world face videos conditioned on head pose estimation," in *CVPR Workshops*. 2012, pp. 130–137, IEEE.

[9] J. Aghajanian and S.J.D. Prince, "Face pose estimation in uncontrolled environments," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2009, pp. 1–11.

[10] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 31, pp. 607–626, 2009.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, 1988.

[13] C. BenAbdelkader, "Robust head pose estimation using supervised manifold learning," in *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, 2010, pp. 518–531.

[14] G. Fanelli, J. Gall, and L. J. Van Gool, "Real time head pose estimation with random regression forests," in *CVPR*. 2011, pp. 617–624, IEEE.

[15] D. Knill and W. Richards, *Perception as Bayesian inference*, Cambridge, 1996.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[17] T. L. Berg A. C. Berg and J. Malik, "Shape matching and object recognition using low distortion correspondences," *In Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[18] J. Kim and K. Grauman, "Boundary preserving dense local regions," *In Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.