Color Prior for Feature-Based 3D Head Tracking

Jixia ZHANG[1], Franck DAVOINE[2], Haibo WANG[1], Chunhong PAN[1]
LIAMA Sino-French IT Laboratory, Beijing, China.
National Laboratory of Pattern Recognition, CASIA[1] CNRS[2]
jixiazhang@gmail.com

Abstract

This paper proposes a novel color prior to reduce outlier keypoints for feature-based 3D head tracking. It is more effective than existing approaches for two reasons. First, the color information of a keypoint is represented by a powerful covariance descriptor (RGB Sigma Set). Second, the color prior is modeled from the posteriors of a discriminative classifier based on Random Projection trees. To demonstrate its strength, we integrate the prior into a 3D head tracking system that estimates pose by matching a set of wide-baseline keypoints. Our experiments show that most outliers are rejected by the proposed prior, which in turn significantly improves the accuracy, speed and robustness of the tracking system.

1 INTRODUCTION

Feature-based tracking via wide-baseline matching has attracted increasing attentions recently for its computational efficiency [8]. One crucial problem with it is how to reject the outlier features in a fast and effective way. Most existing approaches handle this relying on a motion prior derived from previous state or a fixed dynamic model which cannot guarantee to be reliable permanently [8]. In this paper, we propose to utilize a color prior to deal with this problem and verify its efficiency within a 3D head tracking application.

Our motivation to use color clue comes from the observation that it provides more stable and robust information to distinguish inlier and outlier keypoints. Color information has been researched for many years and proven to be effective for face detection and tracking. A good literature survey was given in [10]. The conventional attempts focused on generative models. Unfortunately, none of them is fully insensitive to the various affects, such as illumination and races. It is only recently that discriminative knowledge start to be used for object tracking.

We propose a different approach to use color to reject outliers in feature-based 3D head tracking. Rather than exploring a discriminative color space, we use a discriminative model to distinguish inliers and outliers which relies on the sigma set [3] and random projection trees [2]. Sigma set is a second order statistical descriptor demonstrating good discriminative power. It is proposed to tackle the problem that the similarity measure of covariance matrix is very time-consuming [7]. Random Projection tree (RP tree) is a variant of the K-D tree widely used in machine learning as a spatial partitioning data structure. RP tree has been proposed in [2] as manifold-adaptive spatial data structures for modeling data with a low intrinsic dimensionality and lying artificially in a high D-dimensional space. Its efficiency has been proved in many applications. Forests of RP quantization trees have demonstrated their relevance in [9] for face recognition where each face is represented as a histogram of quantized high dimensional and near invariant features.

Our model describes the points by our named RGB Sigma Set and identifies outliers by RP based classification trees with color information only. Its efficiency is verified through its power to identify outliers and the tracker's improvement by integrating it as a prior.

The structure of the paper is as follows: Section 2 formulates the problem and the creation of color prior is presented in Section 3. Section 4 introduces a feature-based head tracker. Section 5 shows our experimental results and Section 6 closes the paper.

2 Problem formulation

In feature-based head tracking, the key step is to establish a set of geometric 3D-to-2D correspondences by matching individual features to a database of features learnt from reference images. The feature is computed from a local patch around a keypoint in graylevel space. For the i sample $\widetilde{\mathbf{k}}_i$ in our keypoint database $\left\{\widetilde{\mathbf{k}}_1 \dots \widetilde{\mathbf{k}}_m\right\}$, denote $\mathbf{f}(\widetilde{\mathbf{k}}_i)$ as its graylevel feature and $\mathbf{U}_i$ its associated 3D position. At run-time, given the keypoint set $\mathbf{K} = \{\mathbf{k}_1 ... \mathbf{k}_n\}$ from a video frame, each geometric correspondence $\mathbf{U}_i \leftrightarrow \mathbf{v}_j$ is determined by the feature matching with highest probability,

$$\arg \max_{\mathbf{U}_i \leftrightarrow \mathbf{v}_j} P(\mathbf{f}(\widetilde{\mathbf{k}}_i)|\mathbf{f}(\mathbf{k}_j)) \tag{1}$$

where $\mathbf{v}_j$ is the 2D position of $\mathbf{k}_j$. In practice, this matching can cause outlier correspondences from background clutter, mostly because the used feature descriptor is not discriminative enough. If there were excessive erroneous correspondences, it will be impossible to find the consistent pose parameters. In observation, color information is usually distinctive enough to distinguish face from background clutter. Denote $\mathbf{c}(\mathbf{k}_j)$ as the color feature of keypoint $\mathbf{k}_j$. To integrate color, we change the above decision function to be

$$\arg \max_{\mathbf{U}_i \leftrightarrow \mathbf{v}_i} P(\mathbf{f}(\widetilde{\mathbf{k}}_i)|\mathbf{f}(\mathbf{k}_j)) f(\mathbf{c}(\mathbf{k}_j)) \tag{2}$$

where $f(\mathbf{c}(\mathbf{k}_j))$ is a binary function that stands for the inlier/outlier attribute of $\mathbf{k}_j$ given $\mathbf{c}(\mathbf{k}_j)$. Since it gives a *priori* confidence to feature matching from color measurement, we call it color prior. In the following, we will describe how to yield and use this prior.

3 Color prior

The color prior is defined to identify each keypoint as inlier or outlier and its pipeline is illustrated in Fig. 1.

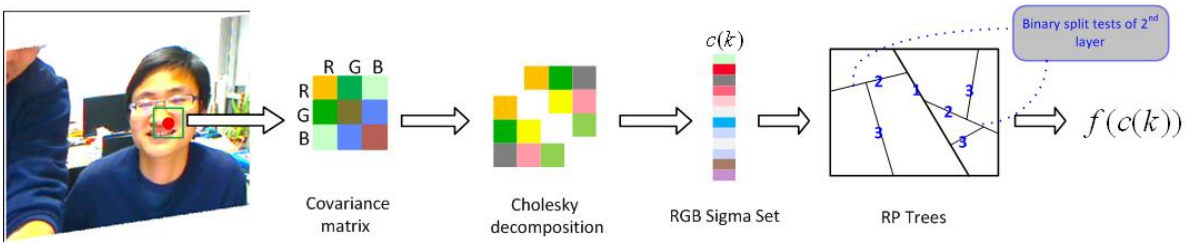

Figure 1. Pipeline of the color prior. The covariance matrix of RGB values in the local patch of keypoint $\mathbf{k}$ is calculated and the RGB sigma set $\mathbf{c}(\mathbf{k})$ is built by decomposing the covariance matrix. Then $\mathbf{c}(\mathbf{k})$ is classified by the RP based classification trees with the numbers illustrating the tree layers. The classification responses are finally used to generate the binary color prior $f(\mathbf{c}(\mathbf{k}))$ of keypoint $\mathbf{k}$.

The color feature $\mathbf{c}(\mathbf{k})$ of each point is represented by a novel RGB Sigma Set vector that captures covariant relationship of color channels. And the binary prior is based on the classification response of the Random Projection(RP) based classifier. Note that $\mathbf{c}$ is short for $\mathbf{c}(\mathbf{k})$ in the following sections.

3.1 RGB Sigma Set

The sigma set is a second order statistics derived from the covariance matrix to solve the distance computation problem of covariance matrix [3]. Here we adopt the RGB information to describe each pixel and use the sigma set to represent one keypoint. A $d_p = 3$ dimensional vector $\mathbf{z}_{ij}, j = 1, 2, \ldots, k$ is used to encode the RGB values of each pixel and the covariance matrix [7] of all k pixels in one patch is given as $\mathbf{C}_i = \frac{1}{k-1}\sum_{j=1}^{k}(\mathbf{z}_{ij} - \mu)(\mathbf{z}_{ij} - \mu)^T$ where μ is the mean vector $\mu = \frac{1}{k}\sum_{j=1}^{k}\mathbf{z}_{ij}$.

Let $\mathbf{C}_i = \mathbf{L}\mathbf{L}^\mathbf{T}$ be the Cholesky decomposition on the covariance matrix. A set of d_p-dimensional sigma points $S = \{\mu + \mathbf{l}_1, \mu + \mathbf{l}_2, \ldots, \mu + \mathbf{l}_{d_p}, \mu, \mu - \mathbf{l}_1, \ldots, \mu - \mathbf{l}_{d_p}\}$ are obtained from the lower triangular matrix $\mathbf{L}$ where $\mathbf{l}_i$ is the α-weighted i^{th} column of $\mathbf{L}$, with $\alpha = \sqrt{k}$ and μ is the mean vector talked beforehand. This sigma set represents the statistics of the patch up to second order, and allows simple similarity measurement on a Euclidean vector space. Then we define the color feature $\mathbf{c}_j$ for one patch as the concatenation of these points which is a $d = d_p(2d_p + 1)$-dimensional vector and use it as the input of the following classifier.

3.2 Random Projection based Classification Trees

RP tree is proposed for vector quantization [2] and the efficiency comes from its particular binary splitting method: instead of dividing a given region into two along coordinate directions at the median, they divide along a random direction.

In our discriminative model, d-dimensional keypoint color feature $\mathbf{c}$ is required to be classified. The label for each keypoint is denoted as y. That is, a binary classifier is needed to identify each $\mathbf{c}$ as belonging to face $(y = +1)$ or background $(y = -1)$. To accomplish that, data at each node are separated by maximizing the entropy based score measure $S_c(L, T) = \frac{2\, I_{c,T}(L)}{H_c(L) + H_T(L)}$.

Here $H_c(L)$ and $H_T(L)$ denotes the classification entropy and split entropy and $I_{c,T}(L)$ denotes the mutual information of the split and the classification. The data at each node is split by a simple linear function given as:

$$\textit{if } \mathbf{b}^T\mathbf{c} + t \leq 0, \textit{ split right; otherwise, split left} \quad (3)$$

where $\mathbf{b}$ is a d-dimensional random projection vector and t is a random scalar threshold.

Training the trees is to select best $\mathbf{b}$ and t for each node. Here $\mathbf{b}$ is chosen from a preselected small dictionary of d-dimensional projection directions. Given the whole training set and projection direction dictionary, the trees are constructed. The recursive tree building algorithm stops when the node receive too few data or when its depth reaches a maximum. Once the forest of N trees is trained, all the leaves encode conditional probabilities for each class. We define $P_j^i(\mathbf{c}|y = +1)$ and $P_j^i(\mathbf{c}|y = -1)$ as the conditional probabilities of face and background respectively at j^{th} leaf of i^{th} tree. Hence, $P_j^i(\mathbf{c}|y = +1) = \frac{N_{ij}^+}{N^+}$ and $P_j^i(\mathbf{c}|y = -1) = \frac{N_{ij}^-}{N^-}$. N_{ij}^+, N_{ij}^- are the number of training samples of face and background at this leaf. While N^+, N^- denote the whole number of face and background training data.

At classification stage, each keypoint represented as $\mathbf{c}$ is dropped down the forest and reaches N leaves. Thus, the final conditional probability of $\mathbf{c}$ is the average of all the conditional probabilities of its reaching leaves $P(\mathbf{c}|y) = \frac{1}{N}\sum_{i=1}^{N} P_{l_j}^i(\mathbf{c}|y)$ where $P_{l_j}^i(\mathbf{c}|y)$ is the conditional probability at the reaching leaf of i^{th} tree for the keypoint. According to Bayes equation $P(y|\mathbf{c}) = \frac{P(\mathbf{c}|y)P(y)}{P(\mathbf{x})}$, we can obtain $P(y|\mathbf{c}) \approx P(\mathbf{c}|y)$ by assuming equal prior.

3.3 Defining the Color Prior

The color prior is defined as a 0-1 function by comparing the two class posteriors $P(y = +1|\mathbf{c})$ and $P(y = -1|\mathbf{c})$, that is:

$$f(\mathbf{c}) = \begin{cases} 1 & P(y = +1|\mathbf{c}) > \alpha P(y = -1|\mathbf{c}) \\ 0 & \textit{otherwise} \end{cases} \quad (4)$$

From this prior, a keypoint $\mathbf{k}$ is identified as an outlier or not before graylevel matching, which saves time in practice. Note that we set $\alpha = 0.9$ to lower the true negative rate as compared to $\alpha = 1.0$.

4 Gray Level Feature-based 3D Head Tracking

The 3D head tracking is performed by matching input gray level features to a database of features from a known target using a fast nearest-neighbor algorithm, and then runing a Perspective-n-Point (PnP) technique to recover the 3D pose specifying these matches. To obtain 3D-to-2D correspondences, a 3D model is constructed from two reference images and a database of keypoint features with their 3D positions is learned from the images. At run-time, each individual feature is matched with the database using fast nearest-neighbor indexing. And a set of 3D-to-2D correspondences can be established to recover the underlying 3D pose. The 64-dimensional Speeded-Up Robust Features (SURF) descriptor is used as the graylevel feature. And the fast nearest-neighbor indexing comes from the implementations in the FLANN library [6].

With only the feature matching, motion jitters are inevitable. Hence, a set of short-baseline correspondences, established by tracking optical flow, are also incorporated with local constancy assumption. For easily associating 3D positions, the optical-flow candidates are selected from two sources - the keypoints recognized at previous frame and a subset randomly drawn from model vertices.

Then the consistent 3D pose can be obtained by simultaneously minimizing the reprojection errors of the two correspondence types

$$\arg\min_{\mathbf{R},\mathbf{T}}\{\sum_{i=1}^{N_f} \parallel \mathbf{K}[\mathbf{R}|\mathbf{T}]\mathbf{U}_i - \mathbf{v}_i \parallel + \sum_{j=1}^{N_s} \parallel \mathbf{K}[\mathbf{R}|\mathbf{T}]\mathbf{U}_j - \mathbf{v}_j \parallel\}, \tag{5}$$

where $\mathbf{K}$ is the known camera intrinsic matrix, $\mathbf{U}_i \leftrightarrow \mathbf{v}_i$ is a pair of feature correspondence and $\mathbf{U}_j \leftrightarrow \mathbf{v}_j$ a pair of optical-flow correspondence. The 3D rotation and translation matrices $\mathbf{R}$ and T are the parameters to recover. One may notice that here we treat the reprojection errors of the two correspondence types equally, but to emphasize the dominant role of the feature type, we adapt N_s as N_f is varied at each frame and always keep $N_s \leq N_f$.

Once we obtain an adequate number of graylevel feature and optical-flow correspondences, an efficient non-iterative PnP algorithm [5] was adopted to generate the pose hypothesis and PROSAC [1] instead of the typical RANSAC algorithm was employed to rank the correspondences just before the hypothesis generation.

5 Experiments and results

The performance of the color prior was evaluated in two aspects. Its discriminative ability was firstly examined by comparing with histogram based descriptions. Then, its practical role of color prior was assessed in 3D head tracking application. First, we present the color histograms we are using.

5.1 Color histograms

To verify the discriminative power of sigma set, different histograms are built based on different color spaces. The color spaces considered in this work are: *RGB* color space, *opponent* color space, *normalized RGB* color space and *transformed* color space [10] and their corresponding histograms are denoted as RGBHist, OppHist, rgHist and TransHist respectively. HSV space has been considered quite often in computer vision. Hence, we also tested the histogram of Hue component which is denoted as HueHist.

5.2 Classification results

The discriminative abilities of all the descriptors were evaluated by performing the classifier on the videos from Boston University [4]. The classifier is trained with one single input image. To enhance its robustness, a set of new views of this image are also generated with randomly drawn affine transformations. In total, 20 trees with a maximum depth of 5 are learned and 200 random projection directions are preselected.

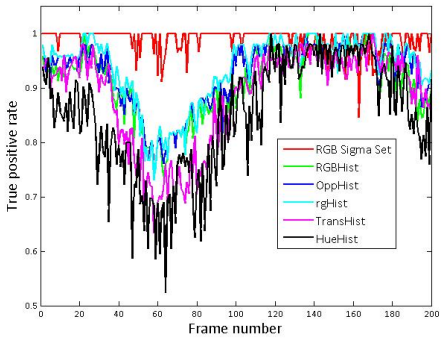

Figure 2. True positive rates of each descriptor on 'ssm1' sequence. Mean TPRs for RGB Sima Set, RGBHist, OppHist, rgHist, TransHist and HueHist are 0.9882, 0.9117, 0.9252, 0.9347, 0.8843, 0.8473 respectively. RGB Sigma Set outperforms others with a near 100% true positive rate.

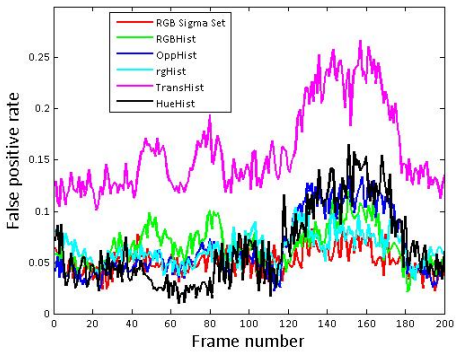

Figure 3. False positive rate of each descriptor on 'ssm1' sequence. RGB Sigma Set leads to a quite low false positive rate.

To measure the performance of our classifier, we manually labeled the ground truth of the 'ssm1' video and computed the true positive rate ($TPR = \frac{TP}{TP+FN}$) of each frame as depicted by Fig. 2. The false positive rates ($FPR = \frac{FP}{FP+TN}$) were also compared to get a more thorough evaluation as shown in Fig. 3. RGB Sigma Set leads to a near 100% TPR and a 0.05% FPR for most frames. Thus, most inlier keypoints are kept

while rejecting most outliers. RGB Sigma Set outperforms other descriptors even with a lower dimensionality because it is a up-to-second-order statistics and is obtained from the covariance matrix of colors. The covariant relationship between color channels leads to the better discriminative power.

We also test the RGBSIFT descriptor which is the combination of SIFT descriptor on each color channel. However, the 384-dimensional descriptor can not discriminate face keypoints from background keypoints well and its mean precision is only 0.628. RGB Sigma Set and color histograms show reliable discriminative ability while RGBSIFT results in a worse classification performance. RGB Sigma Set and color histograms capture statistical color information but RGBSIFT combines color and local shape information.

5.3 Application to head tracking

This subsection assesses the roles of the proposed color prior in 3D head pose tracking. Both quantitative and qualitative experiments were conducted.

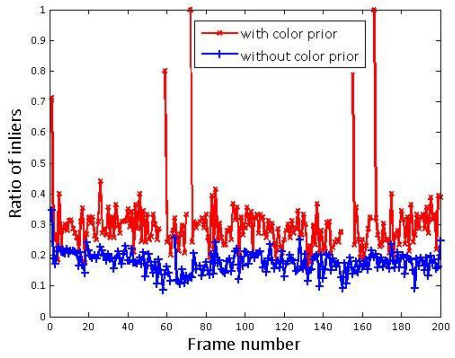

Figure 4. Ratio of inliers for each frame with and without color prior.

The quantitative experiment was conducted with the same 'ssm1' sequence from Boston University [4]. The measure is defined as the ratio of inliers to all the matched keypoints during pose tracking. This is reasonable since it is critical to the final PROSAC selector. Fig. 4 depicts the improvement of ratio of inlier by about 0.2 after the color prior is imposed. Consequently, this improvement saves the number of PROSAC iterations such that pose tracking is speeded up by about two times.

The qualitative experiments were carried out with live-captured challenging video sequences. Fig. 5 shows the comparative results with and without color prior. With the color priors, a larger portion of inliers is reached and the tracker is more robust to large head movements and sudden changes of lighting conditions.

6 Conclusion

Our proposed color prior helps to decide if a keypoint represented by our named RGB Sigma Set is an inlier or not. Its discriminative ability is achieved by Random Projection based classification trees. It is demonstrated to be robust to pose variations, facial expressions and illumination changes. We also verify its merit through a 3D feature based head tracker. We believe that it will also be a good prior for other head-related tasks, such as face localization and face recognition.

References

[1] Ondrej Chum and Jiri Matas. Matching with prosac - progressive sample consensus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 220–226, 2005.

[2] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *Neural Information Processing Systems*, 2007.

[3] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: A small second order statistical region descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1802–1809, 2009.

[4] Stan Sclaroff Marco La Cascia and Vassilis Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 322–336, 2000.

[5] Francesc Moreno-Noguer, Vincent Lepetit, and Pascal Fua. Accurate non-iterative o(n) solution to the pnp problem. In *ICCV'07*, volume 0, pages 1–8, 2007.

[6] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP'09*, pages 331–340, 2009.

[7] O. Tuzel, F. Porikli, and P. Meer. Region covariance: a fast descriptor for detection and classification. In *European Conference of Computer Vision*, pages 589–600, 2006.

[8] Qiang Wang, Wei Zhang, Xiaoou Tang, and Heung-Yeung Shum. Real-time bayesian 3-D pose tracking. *IEEE Trans. on Circuits and Systems for Video Technology*, 16(12):1533–1541, 2006.

[9] J. Wright and G. Hua. Implicit elastic matching with random projections for pose-variant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[10] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:34–58, 2002.

Figure 5. Snapshots of tracking results with color prior (left) and without color prior (right).