User-Independent Face Landmark Detection and Tracking for Spatial AR Interaction

Youngkyoon Jang[1], Eunah Jung[2], Sung Sil Kim[3], Jeongmin Yu[1], and Woontack Woo[1,3(✉)]

[1] CTRI & AHRC, KAIST, Daejeon, South Korea
{y.jang,jmyu119,wwoo}@kaist.ac.kr
[2] School of Computing, KAIST, Daejeon, South Korea
514ah@kaist.ac.kr
[3] GSCT, KAIST, Daejeon, South Korea
mania@kaist.ac.kr

Abstract. We present novel face landmark detection and tracking methods which are independent of user facial differences in a scenario of Spatial Augmented Reality (SAR) interaction. The proposed methods do not require a preliminary general face model to detect or track landmarks. Our contributions include: (i) fast face landmark detection, which is achieved based on our modified Latent Regression Forest (LRF) and (ii) model-independent facial landmark tracking by revising outliers based on a direction and displacement of neighboring landmarks. We also discuss (iii) feature enhancements based on RGB and depth images for supporting several interaction scenarios in SAR environments. We anticipate that the proposed methods promise several interesting scenarios, even under severe head orientation in SAR interaction without wearing any wearable devices.

Keywords: Face landmark detection · Face landmark tracking · Random forest · Virtual reality · Computer vision

1 Introduction

Spatial Augmented Reality (SAR), such as IllumiRoom [9] and RoomAlive [8], provides immersive user experiences by projecting a VR scene onto the room space and expanding an interactive space. However, a user is required to touch a point of the wall in order to interact with the projected virtual object in the SAR environments. Because it is hard to estimate head position and orientation without wearing HMD, estimating head pose, detecting and tracking facial landmarks provides various interactive clues which are available for supporting a more intuitive interaction in SAR environment. Herein, head orientation indicates the direction of the user's view. Moreover, user-independent face landmark detection and tracking is the first step for supporting different users of SAR environment. Thus, without wearing a cumbersome Head-Mounted-Display (HMD), such as HoloLens [1] and Oculus [2], an immersive experience is achievable based

© Springer International Publishing Switzerland 2016
N. Streitz and P. Markopoulos (Eds.): DAPI 2016, LNCS 9749, pp. 210–220, 2016.
DOI: 10.1007/978-3-319-39862-4_20

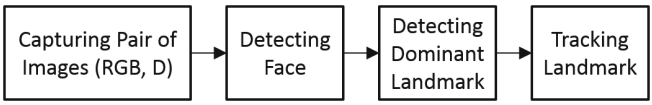

Fig. 1. Block diagram of the proposed framework.

on the User-independent face landmark detection and tracking in an indoor environment.

Main contributions of our proposed method include:

(1) Fast head orientation estimation: Our modified Latent Regression Forest for face landmark detection guarantees very fast detection performance. Moreover, the proposed method detects face landmarks independent of user facial differences.

(2) Model-independent facial landmark tracking: Based on the detected landmarks as an input, the proposed method could track the landmarks along the video sequences, which is independent of user's facial expression changes. Because the proposed method does not require a preliminary model for a general face, it provides more accurate tracking performance.

(3) Discussions of novel feature enhancement based on RGB-D images: For improving the accuracies of landmark detection and tracking, we discuss a novel type of feature configuration which utilizes the concept of Local Angle Pattern [7].

2 Methodology

2.1 Overview

The proposed method is as shown in Fig. 1. At first, RGB-D camera captures a pair of images including synchronized color and depth images. Based on a depth image, then, our method detects the face region of interests. After that, the center coordinate of the detected face region is transformed into the coordinate of the color image. Then, by utilizing a cropped face image as an input, our proposed method detects multiple dominant facial landmarks in a coarse manner. For that, we adopted and modified the Latent Regression Forest [13] for targeting face modality. The coarsely detected dominant landmarks work to specify the searching space for specifically aligned landmark detection. Finally, our proposed method tracks the landmarks based on the proposed outlier rejection methods.

2.2 Region of Faces Detection

Our face detection is achieved by utilizing J. Shotton [12]'s body joint estimation method. By taking the two pixel test, which is based on the normalized offsets to be calculated, we only trained face and background classes. Based on

the coarsely detected region, we redefine a searching space and then did per-pixel classification for more precise face region detection. Based on the detected face region, a cropped and normal-sized face image is used as an input for face landmark detection.

2.3 Dominant Landmark Detection

We use Latent Random Forest (LRF) [13] for face dominant landmarks. The LRF utilizes a face landmark topology to keep a relative position structure of landmarks. We designed a face landmark topology to guide landmark detection (Sect. 2.3-A), built LRF following the designed topology (Sect. 2.3-B), and designed testing procedure using learnt LRF (Sect. 2.3-C)

A. Face Landmark Topology. To enhance the landmark detection process, we utilize a hierarchical context of dominant landmarks based on a topology of face landmarks. Given ten dominant landmarks in Fig. 2, a face landmark topology has a binary tree structure. From the center position of the face image represented as a root node, we could reach to every landmark stored at the leaf nodes of topology. Each node in the topology has its two children which have the subset of its parents' landmark set, respectively. When it reaches the node that has only one dominant landmark in the subset, we define the node as a leaf node (See Fig. 2).

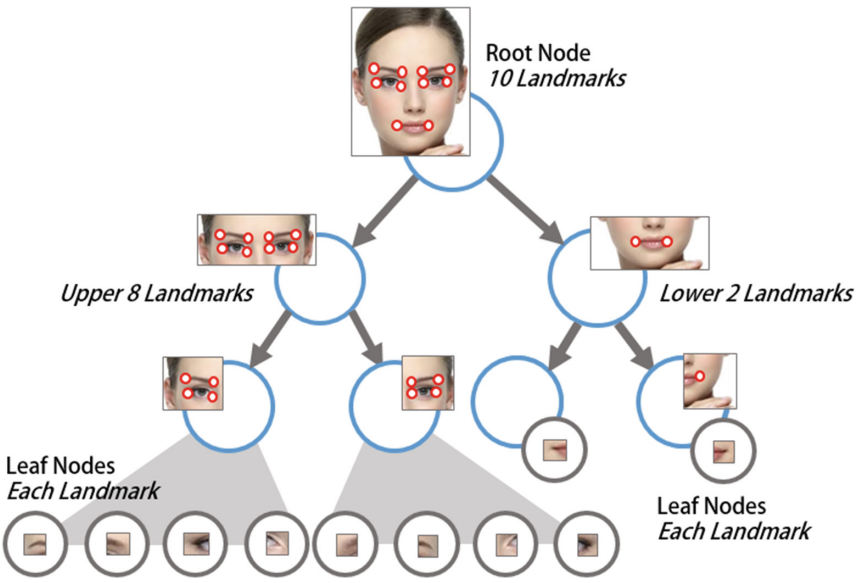

Fig. 2. Face landmark topology model.

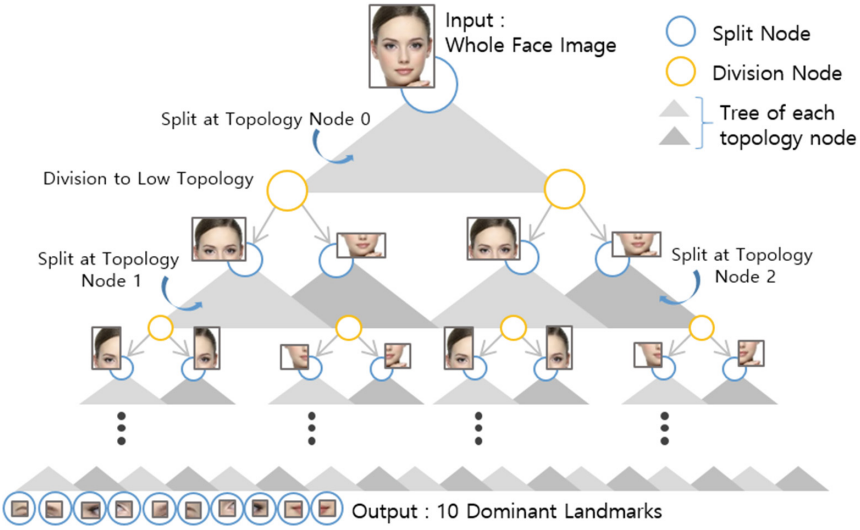

Fig. 3. Latent regression tree.

B. Learning Latent Regression Forest (LRF). A Latent Regression Forest (LRF) [13] is an ensemble of random decision trees, which is originally proposed to estimate articulated hand posture. In this paper, we adopted the LRF to estimate the facial landmarks, searching from coarse level to fine level based on the designed face topology model. Learning LRF is performed in a divide-and-conquer way by taking the whole face image and ground-truth landmarks as input and ending with each landmark detected as output. Each node of a decision tree in LRF is set by one of three types: split, division, and leaf. Split nodes function to split the training dataset to two subsets by the split function. Division nodes divide the scope of facial landmark set according to the face topology. When the scope of facial landmark includes only one landmark, it terminates by storing the relative landmark position at the leaf node.

Given our face topology model M, for each node $i \in M, i = 0, \cdots, |M|$, it has parent node $p(i)$ and its child nodes $l(i)$ and $r(i)$. For each training RGB face image I, we define ${\rho_i}^I$, the center position of a landmark set corresponded with each topology node i. Each latent regression tree is trained corresponding with each topology stage. A root node of LRT takes the whole scope of facial landmarks according to a root node $i = 0$ of the topology model. As the tree grows, it separates the scope of landmarks according to $l(i)$ and $r(i)$ until the topology reaches the leaf node. At each node, we split the training data S into two subsets S^l and S^r by the split function f_i and threshold τ_i randomly generated. The learning is proceeded under the context of a topology node i. A split function f_i and the subsets are defined as:

$$f_i = I({\rho_i}^I + u) - I({\rho_i}^I + v), \tag{1}$$

$$S^l = \{I | f_i(I) < \tau_i\}, S^r = S \backslash S^l, \tag{2}$$

where $I(\cdot)$ is the pixel value of certain location, vectors u and v are random normalized offsets. We set the split function f_i which shows the largest information gain value, while if the information gain value could not improve from the previous node step, the learning process enters the division step. The information gain under the context of a topology node i is defined like [13] as:

$$IG_i(S) = \sum_{m}^{l(i),r(i)} tr\Big(\sum_{im} S\Big) - \sum_{k}^{l,r} \frac{S^k}{|S|}\Big(\sum_{m}^{l(i),r(i)} tr\Big(\sum_{im} S^k\Big)\Big), \tag{3}$$

where $\sum_{im} \chi$ is the sample covariance matrix of the set of offset vectors $\{(\rho_m^I - \rho_i^I) | I \in \chi\}$. The offset vectors are the offsets from the current center position to each center of two subsets.

Given the training data which are face images, at division step, each data is divided by the center of the selected offset vectors. Its children nodes process its own learning on a finer scope of the training data. (See Fig. 3) The offset vectors $\theta_m = (\rho_m^I - \rho_i^I), m \in \{l(i), r(i)\}$ are stored in the division node.

Split and division process are repeated until a corresponding topology node is the leaf node of the topology which represents one final landmark. At each leaf node, we save the offset vectors from the center of its parent node to the landmark.

C. Testing. Given a detected face image as an input, it goes into each Latent Regression Tree in LRF, starting from the center of the face image with the root node of a topology. At each split node, the test image is checked with the split function saved in the node, traversing to the left side or the right side and repeats the process until reaching at division node. At each division node, the face image is divided into two sub-regions according to the children nodes of the current node in the topology and the landmark position is accumulated with the offset vectors saved in the division node. When reaching a leaf node and accumulating the offset vectors, all dominant landmark positions can be estimated.

2.4 Model-Independent Landmark Tracking

In this section, we present a model-independent landmark tracking method. Recently, model-based methods [4,11,15] have been popularly used for landmark tracking and have achieved promising tracking results. However, these methods are not suitable for tracking various face appearances, and their tracking performances heavily rely on a number of training samples and optimization methods. To overcome these problem, we propose a model-independent landmark tracking method which is based on dense optical flow [15] and the displacement of neighborhood landmark information. Specifically, after detecting the dominant landmarks, each detected landmark $p_t^l = (x_t^l, y_t^l)$ at frame t is tracked to the next frame $t+1$ using the median filtering kernel M in a dense optical flow field $G = (u_t, v_t)$.

$$p_{t+1}^l = (x_{t+1}^l, y_{t+1}^l) = (x_t^l, y_t^l) + (M \cdot G)|(\bar{x}_t^l, \bar{y}_t^l), \tag{4}$$

where $\cdot$ is the convolution operator, and $(\bar{x}_t^l, \bar{y}_t^l)$ is the rounded position of (x_t^l, y_t^l).

During landmark tracking, landmarks tend to drift from their previous locations due to the abrupt fast motions of the face. To revise such outlier landmarks, we use displacement information of neighboring landmarks, which is illustrated in Fig. 4. In detail, we first define the neighboring landmarks by considering their geometric information and the partial components of the face. Then, the outlier landmarks are modified by their neighboring displacements and the mean of their directions.

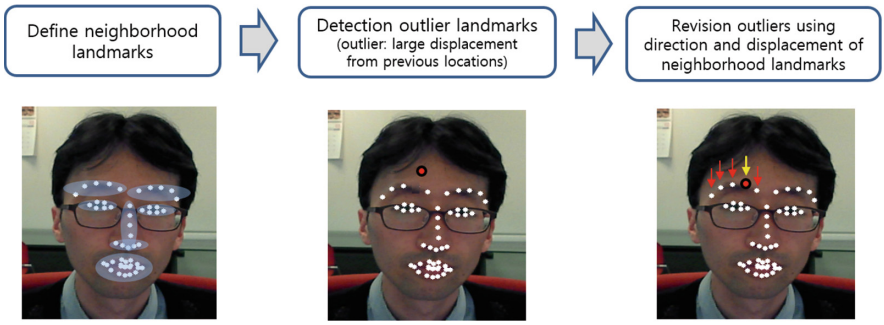

Fig. 4. Process of revision of outlier landmarks.

When an overall landmark tracking error is larger than a threshold, we reinitialize the landmarks positions using the proposed dominant landmark detection method.

3 Implementation

3.1 LRF-Based Landmark Detection

To perform the dominant landmark detection method, each training sample consists of a cropped face image and a ground truth of ten dominant facial landmarks. Because a LRF is trained based on pixel value of the face image, we normalize each face image to a fixed-size image (30×30 pixels in our implementation) and make a feature vector which is 900 pixel values of a normalized image. At the division step, we mark invalid pixel values as -1 in each feature vector as the facial landmarks are divided to two subsets. A two-pixel difference test proceeds on valid elements of feature vectors. For testing landmarks, when traversing the trees and accumulating the offset vectors which are saved in division nodes, we calculate the offset values by using the voting mechanism. This is found to be more accurate than averaging all offset values.

(a) (b)

Fig. 5. Results of face landmark tracking: (a) tracking result of KLT tracker (b) tracking result of our proposed method.

3.2 Model-Independent Landmark Tracking

For testing the proposed tracking method, we first capture face motion clips which contain in-plane rotation of face from a commodity RGB camera. Then, we conduct an experiment using the proposed method and the other model-independent landmark tracking method (namely KLT tracker) [11] which is based on sparse optical flow information. Figure 5 shows their tracking results.

As shown in Fig. 5, our proposed method (Fig. 5(b)) outperforms KTL landmark tracker (Fig. 5(a)) with respect to the in-plane rotation of face situation. From the experiment, we confirm the feasibility of the proposed landmark tracking method which does not use a learned face model.

4 Discussion

So far, we have tackled user-independent landmark detection and tracking methods. However, it is still processed based on the images captured from frontal viewpoint cameras. Thus, in order to enhance the features, which can be used for the scenarios of face rotation in SAR environment, we discuss feature enhancement method in this section.

Defining RGB-D feature is completed as shown in Fig. 6. The color and depth image acquired from a camera is processed parallely into local binary pattern (LBP) image and surface normal image. The color image is perceived as a matrix of RGB pixels by a camera in order to apply Local Binary Pattern (LBP) [3]. LBP operator is one of the most efficient and effective image features, frequently used in face recognition and detection. Despite its advantages, however, LBP features still suffer in terms of robustness in situations where instant change of luminosity or face orientation occurs. To overcome this challenge, the proposed feature integrates LBP feature with depth data.

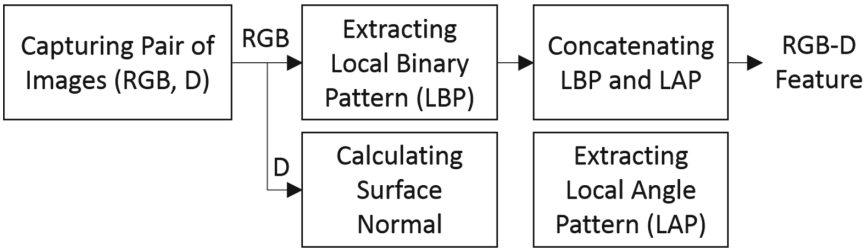

Fig. 6. RGB-D feature definition flow diagram.

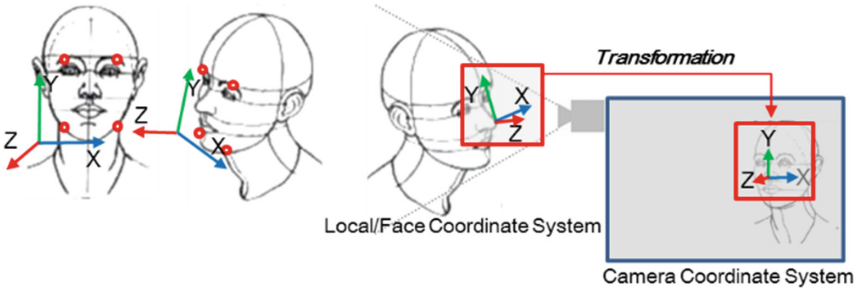

Fig. 7. Acquiring transformation matrix between local/face coordinate and camera coordinate based on dominant facial landmarks.

Along with color images, a stream of depth images is captured from TOF camera simultaneously, and these depth images represent the distance between objects and the camera. The proposed method utilizes the depth images by calculating the surface normal of each pixel by finding a vector for each pixels which is orthogonal to the plane.

In the process of generating our proposed Local Angle Pattern, a coordinate transformation is done in order to justify the orientation difference between camera and local/face surface normal. The original surface normal vectors are adjusted by applying a transformation matrix resolved from the relationship of four dominant facial landmarks in camera coordinate and local/face coordinate, as shown in Fig. 7.

Transformation Matrix Tr used in coordinate adjustment is defined by the rotation and translation, and it is referenced to transform 3D vectors from camera coordinates L to local coordinates L', as shown in Eq. 5.

$$\begin{pmatrix} L'_{1x} & L'_{1y} & L'_{1z} \\ L'_{2x} & L'_{2y} & L'_{2z} \\ L'_{3x} & L'_{3y} & L'_{3z} \\ L'_{4x} & L'_{4y} & L'_{4z} \end{pmatrix} = Tr \begin{pmatrix} L_{1x} & L_{1y} & L_{1z} \\ L_{2x} & L_{2y} & L_{2z} \\ L_{3x} & L_{3y} & L_{3z} \\ L_{4x} & L_{4y} & L_{4z} \end{pmatrix}, \tag{5}$$

Transformation Matrix Tr consists of 3 variables: internal calibration matrix A, rotation matrix R, and translation vector T as shown in Eq. 6. The

inverse matrix of rotation matrix R inside of transformation matrix Tr is applied on camera coordinates surface normal SN, such that consistent local surface normal SN' image is generated under various face rotation, as shown in Eq. 7.

$$Tr = A[R|T], \tag{6}$$

$$\begin{pmatrix} SN_{1x} & SN_{1y} & SN_{1z} \\ & \vdots & \\ SN_{nx} & SN_{ny} & SN_{nz} \end{pmatrix} [R]^{-1} = \begin{pmatrix} SN'_{1x} & SN'_{1y} & SN'_{1z} \\ & \vdots & \\ SN'_{nx} & SN'_{ny} & SN'_{nz} \end{pmatrix}, \tag{7}$$

Before integrating with the LBP feature to form a RGB-D feature, the transformed surface normal are encoded in 8-bit code by applying locality principle [6], as shown in Fig. 8.

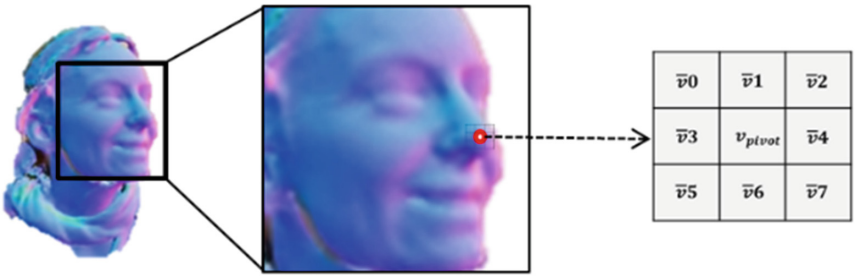

Fig. 8. Applying locality principle to surface normal image.

For each pixel, the inner product of pivot and its neighboring surface normal vector are calculated. If the inner product of two vectors is greater than the threshold, the feature concatenates 1 and if not, 0. As a result, each pixel produces an 8-byte local angular pattern (LAP), as defined by the following equation:

$$\theta_i = cos^{-1}\Big(\frac{dot(v_{pivot}, v_i)}{\|v_{pivot}\| \cdot \|v_i\|}\Big), \tag{8}$$

$$lap[i] = \begin{cases} 0, & \text{if } \theta_i < \text{threshold.} \\ 1, & \text{if } \theta_i \geq \text{threshold.} \end{cases}, \tag{9}$$

$$LAP(x, y) = (lap[0], lap[1], \cdots, lap[7]), \tag{10}$$

The LAP feature defined here is integrated with LBP to form RGB-D feature, and it is used for both holistic and patch-based face detection and tracking.

The proposed RGB-D feature has the merit of supplementing the limitations of color-based features, such as SIFT [10], HOG [5], Viola-Jones [14], or LBP [10] itself. Many of the above methods showed that they rely on edge/corner extraction for detecting faces, which becomes easily vulnerable in rotation, translation,

or slight deformation. Also, upon immediate centralization of image contrast, losses of feature detection occur, thus being unstable for SAR interaction.

Our proposed RGB-D feature, on the other hand, utilizes both color and depth data and proved to be more robust under rapid change of light conditions and face rotation. The fact that LAP [7] is not affected by light conditions not only improves the detection accuracy in various light conditions, but also stabilizes the tracking in dynamic head orientation by generating features based on face/local coordinates. Therefore, even under conditions where color images are suddenly saturated such that no RGB features are extractable, the LAP becomes the reference to where the facial landmarks are located.

5 Conclusions

This paper presents user-independent face landmark detection and tracking methods for spatial AR interaction scenarios. In spatial AR interaction environments, a user has to approach and touch the projected object on the wall in order to interact with virtual objects, which is cumbersome. To detect the facial landmarks, we adopted and modified Latent Regression Forest (LRF), specifically for face modality. Because it utilizes a face image to detect all the landmarks of a face, it is fast and invariant to users' facial differences. In addition, to track the landmarks, we proposed a model-independent tracking method. Finally, we discussed feature enhancement method, which could be used for both detection and tracking. We expect our proposed user-independent facial landmark detection and tracking methods would be useful in SAR interaction scenario.

Acknowledgments. This work was supported by DMC R&D Center of Samsung Electronics Co.

References

1. Microsoft HoloLens. https://www.microsoft.com/microsoft-hololens/. Accessed 25 Sept 2015
2. Oculus, V.R.: https://www.oculus.com/. Accessed 25 Sept 2015
3. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
4. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. ACM Trans. Graph. **32**(4), 41: 1–41: 10 (2013)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) International Conference on Computer Vision & Pattern Recognition, vol. 2, pp. 886–893. INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334. June 2005. http://lear.inrialpes.fr/pubs/2005/DT05
6. Denning, P.J.: The locality principle. Commun. ACM **48**(7), 19–24 (2005)
7. Jang, Y., Woo, W.: Local feature descriptors for 3d object recognition in ubiquitous virtual reality. In: 2012 International Symposium on Ubiquitous Virtual Reality, Daejeon, Korea (South), 22–25 August 2012, pp. 42–45 (2012)

8. Jones, B., Sodhi, R., Murdock, M., Mehra, R., Benko, H., Wilson, A., Ofek, E., MacIntyre, B., Raghuvanshi, N., Shapira, L.: Roomalive: magical experiences enabled by scalable, adaptive projector-camera units. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, UIST 2014, NY, USA, pp. 637–644. ACM, New York (2014)
9. Jones, B.R., Benko, H., Ofek, E., Wilson, A.D.: Illumiroom: peripheral projected illusions for interactive experiences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, NY, USA, pp. 869–878 (2013). http://doi.acm.org/10.1145/2470654.2466112
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
11. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI 1981, vol. 2, pp. 674–679. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1981)
12. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A.: Efficient human pose estimation from single depth images. In: Transaction on PAMI (2012)
13. Tang, D., Chang, H.J., Tejani, A., Kim, T.K.: Latent regression forest: structured estimation of 3D articulated hand posture. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I-511–I-518 (2001)
15. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition CVPR 2011, pp. 3169–3176. IEEE Computer Society, Washington, DC, USA (2011)