Perception-Based Design for Tele-presence

Santanu Chaudhury[1], Shantanu Ghosh[1,2], Amrita Basu[3], Brejesh Lall[1], Sumantra Dutta Roy[1], Lopamudra Choudhury[3], R. Prashanth[1], Ashish Singh[1], and Amit Maniyar[1]

[1] Multimedia Lab, Department of Electrical Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi - 110 016, India
schaudhury@gmail.com,brejesh@ee.iitd.ac.in, sumantra@ee.iitd.ac.in,prahanth.prbest@gmail.com, ashish.iitd07@gmail.com,maniyar.amit@gmail.com

[2] Behavioural & Cognitive Science Lab, Department of Humanities & Social Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi - 110 016, India
sghosh.neu@gmail.com

[3] School of Cognitive Sciences, Jadavpur University, Kolkata - 700 032, India
amrita8@gmail.com, choudhuryl@yahoo.com

Abstract. We present a novel perception-driven approach to low-cost tele-presence systems, to support immersive experience in continuity between projected video and conferencing room. We use geometry and spectral correction to impart for perceptual continuity to the whole scene. The geometric correction comes from a learning-based approach to identifying horizontal and vertical surfaces. Our method redraws the projected video to match its vanishing point with that of the conference room in which it is projected. We quantify intuitive concepts such as the depth-of-field using a Gabor filter analysis of overall images of the conference room. We equalise spectral features across the projected video and the conference room, for spectral continuity between the two.

Keywords: Tele-presence, Homography, Texture Transfer, Perceptual Continuity, Spectral Correction.

1 Introduction

This paper presents a perception-driven approach to low-cost tele-presence systems. An immersive experience to a tele-presence system (so that the projected video of another room appears to be 'continuous' with regard to the people in the room, for instance) can come with 'normalisation' of two parameters: the geometry, and the spectral characteristics, in the projected video. We present an approach driven by perceptual cues and cognitive results, to enable the same in a low-cost video conferencing system. To the best of our knowledge, such an approach has not been reported before in the literature.

A teleconferencing system represents an elaborate set of projection systems that allow rendering of 3-D objects from a real environment in specific 2-D perspectives on a projective surface with a real environment depending on the

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 154–159, 2011.

projection angle of the camera. Such a technologically mediated communication involves certain degree of 'presence' or a 'feeling of being there' that allows the conferees to respond realistically to events and situations in the projected to real environment. Perception of such extensions may occur due to geometrical alignments, rendering of objects, and illumination patterns between real and projected environments. The present 'state of the art' teleconferencing systems utilize very high-end studios congruent in geometrical alignments and illumination patterns to enhance this 'same-room effect'. However, providing exactly similar environments for the virtual and the real counterparts is often difficult to achieve and may not provide a 'seamless' integration of the real and the virtual environments. There are obvious constraints associated with such a system, like small size and relative independence of location of the projection screen that drives the need for designing a cognitive cue-based geometrical and spectral correction algorithm for an 'immersive' experience during conferences. Hence we favour a scene recognition approach to the problem over the current reconstruction techniques and introduce an algorithm that leads to an enhanced visual continuity. This is achieved by extracting semantic category information and visual congruence.

In this paper, we present the results of a computational approach based on classification and organisation of tele-presence video conferencing frames into congruent and non-congruent domains with respect to projected-to-real symmetry-based continuity and depth perception of the two scenes. We utilise lower level information such as textures and perceptual feedback to build up a confidence contour of the context and compute the perceptual difference between the two contexts. Further, our method minimizes the perceptual difference by adjusting the intermediate level properties used for identification of contexts. The perceptual cue used is the continuity of vanishing lines from the projected images, matching the vanishing lines in an image of the conference room. Our method scores on three points: determination of the spectral match between real-to-projected environments, a continuous organisation between the real and the projected environments and determination of a geometric alignment match between the projected and real scenes.

2 Tele-presence and Video Conferencing

A scene may be described with reference to the observer who has a 'fixated point of view' in an image. Oliva and Torralba [5] propose to consider the absolute distance between the observer and the fixated zone as a principal property to define a 'scene'. A 'view on a scene' begins when there is actually a larger space between the observer and the fixated point, usually after 5 meters (typically it refers to a single image in image processing/computational vision). In a video conferencing situation however, an ambiguity in the perception of the fixated point is created by introducing a projected image (thereby creating at least a two scene system). Perception of a video conferencing scene may therefore be defined as a composite of two scenes where the fixation point is altered by

an introduction of a projected scene. We seek to create a 'unified' scene in two different ways. First, we try to align the geometry of the two scenes in a computationally simple method (Section 3). Next, we seek to unify the spectral characteristics of the projected video, and a conference room, using low-level features which are descriptors of a higher level percept (Section 4). This creates a visual illusion of the same room in such a composite scene by describing feature composites to yield higher level features. The human visual system uses this for fast and efficient detection of scene congruence between the projected and real environments.

3 Geometric Correction: Novel View Generation

The first step in our approach is to correct for the geometry of the projected video scene, so that it appears to be perceptually 'continuous' with the geometry of the room in which the video is projected. There is a wealth of literature on novel view generation, given either a number of views, or a single one. (Representative references for the two are [2] and [1], respectively). For multiple views, for a slowly moving camera, one can establish correspondences between different views of the same scene (using any tracker, for instance). A typical approach first computes the projective structure, and using further constraints, updates it to affine, metric or Euclidean, depending on the nature of the constraints. For the single view method, again one needs specific constraints [1]. These methods are computationally complex, and often quite cumbersome.

We adopt a different methodology - one based on the ideas of Hoiem, Efros and Hebert [3]. This is a fully automatic method for creating a 3-D model from a single photograph. We use their basic idea of trying to statistically model geometric classes defined by their orientations, rather than trying to recover the precise geometry. In our case, we learn labels corresponding to coarse categories 'horizontal', 'vertical', and 'background' from representative images of video conferencing scenes. Based on the statistical learning of the labels, given a new video conferencing scene, the algorithm segments image regions, groups them together, and then uses the learnt information into the three categories, and then attempts to recreate the structure of the scene for the three categories. Given the estimated structure for the horizontal and vertical segments for instance, we specify sample camera parameters for the new viewpoint. The new viewpoint is estimated according to the orientation of the lines in an image of the video conferencing environment. We adopt a Hough transform-like approach, and vote on the different possible vanishing points (otherwise known as the FoE: 'focus of expansion') within the image of the video conferencing room in question. We assume that the one getting the highest vote is the required point, and we orient the camera parameters in order to project the above estimated structure on the projection screen in the video conferencing room, in order to equalise the two vanishing points. This is the essence of trying to 'equalise' the two viewing geometries. Fig. 1 shows two examples of using such an automatic correction for the projected frames. In the first one, the desired image has the

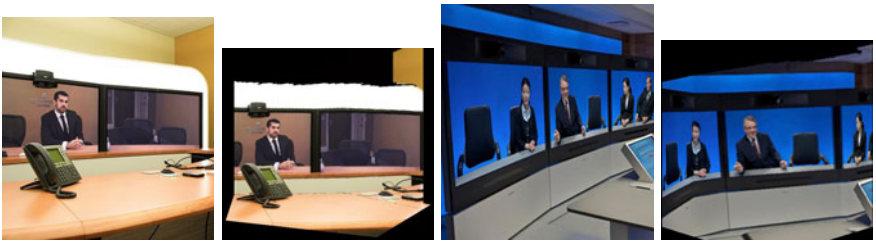

Fig. 1. Geometric Correction for two cases: each case shows the uncorrected, and the 'corrected' image, side-by-side. Details in Section 3.

principal vanishing point somewhere towards the centre of the image. This is slightly to the left, for the second example. The algorithm performs a texture mapping for each planar surface, and pastes this on the new projected orientation of the planar surface, using a homography between the planes in the two - the original image, and the projected image in the new orientation.

The above method - while being simple and easy to compute, suffers from obvious generalisation limitations. It works well for planar surfaces whose overall orientations are estimated well. For complex objects and non-planar ones such as human beings, the method fares poorly. For such cases, we use a simple heuristic. We can extract moving objects such as human beings using a simple but robust motion estimation algorithm, such as [4]. This algorithm segments out any number of foreground objects relatively quickly, against a slowly moving or static background, which is the case in any video conferencing application. The planar regions can be placed at their re-oriented locations, and the other objects pasted on top. The 'background' labelled regions are projected using a texture map homography.

The geometric alignment achieved in the previous section renders video frames, where the scene is conceived of as an 'image-within-an-image'. This alignment also allows us to 'equalise' the spectral parameters in the foreground and background of this composite scene, which completes the illusion of perceived continuity. This gives the entire scene a cognitively congruent interpretation.

4 Spectral Correction

We perform spectral correction between the given conference room, and the video that is projected in it. We take images of the conference room, and consider the projected video that has been subjected to the geometric correction of the previous section. We consider images in a 'perceptual colour space' such as the HSI (Hue-Saturation-Intensity) model. Initial results of our perceptual continuity experiments suggest that users get a better immersive experience with regard to the following three parameters. First, it is easy to have an overall hue of yellow in the conference room. We specify the histogram in the hue parameter

in the projected video, to peak around yellow. Next, we perform a histogram equalisation on the saturation and intensity components.

We described the organisation of the projected-to-real tele-presence scenes along two semantic axes providing an ideal Spectral and Geometric Matching Template (SGMT) (i.e., from general indoor scenes to specific video conferencing scenes; from highly textured 'indoor'-scapes to no textures, from 'deep' rooms to 'shallow' rooms and from 'artificial' to 'natural' indoor scenes with respect to an alignment of vanishing lines). All images were pre-processed to reduce the effects of large shadows that may hide important parts of the scene and to minimize the impact of high contrasted objects which would disturb the power spectrum shape of the background image. First, we apply a logarithmic function to the intensity distribution. Then we attenuate the very low spatial frequencies by applying a high pass filter. Next, we apply an adjustment of the local standard deviation at each pixel of the image. This makes large regions of the image being equally bright.

To create the semantic axes, we choose two sets of prototypical scenes that determine the two extremities of the proposed semantic axes: 'shallow' – 'deep', and 'continuous' – 'discontinuous'. We extract spectral parameters (quantised PCA values from the output of a Gabor filter bank: details below) from the exemplar images, and use them to numerically quantify the extremities. A discriminant analysis computes the axes that both maximises the distance between the two prototypical groups and minimises the standard deviation of the images belonging to the same group.

The transfer function of each Gabor filter is tuned to spatial frequency f_r in the direction determined by the angle θ:

$$G(f_x, f_y) = K \exp(-2\pi^2(\sigma_x^2(f'_x - f_r)^2 + \sigma_y^2)f'^2_y) \tag{1}$$

where f'_x and f'_y are obtained by rotation of the spatial frequencies $f'_x = f_x \cos\theta + f_y \sin\theta$ and $f'_y = -f_x \sin\theta + f_y \cos\theta$. σ_x and σ_y give the shape and frequency resolution of the Gabor filter. K is a constant. The full set of filters is obtained by rotation and scaling of this expression. This gives a high frequency resolution at low spatial frequencies and a low frequency resolution at high spatial frequencies. The values σ_x and σ_y are chosen in order to have coincidence in the contour section of the magnitude at -3 dB. Given an image, its semantic content is invariant with respect to a horizontal mirror transformation of the image. Therefore, we compute the symmetric energy outputs of the Gabor filters which are invariant with respect to a horizontal mirror transformation:

$$\Gamma_{f_{r,\theta}} = \int\int |I(f_x, f_y)|^2 [G^2_{f_{r,\theta}}(f_x, f_y) + G^2_{f_{r,\pi-\theta}}(f_x, f_y)] df_x df_y \tag{2}$$

where $|I(f_x, f_y)|^2$ is the power spectrum of the image, $G_{f_r,\theta}$ and $G_{f_r,\pi-\theta}$ are two Gabor filters tuned to the spatial frequencies given by the radial frequency f_r and the directions θ and $\pi - \theta$. We then use normalised features:

$$\tilde{\Gamma}_{f_{r,\theta}} = (\Gamma_{f_{r,\theta}} - E(\Gamma_{f_{r,\theta}}))/std(\Gamma_{f_{r,\theta}}) \tag{3}$$

where E and std are the mean and the standard deviation of the features $\Gamma_{f_{r,\theta}}$ computed over the entire image database. For each image, we have a feature vector defined by the collection of $\tilde{\Gamma}_{f_{r,\theta}}$ obtained at different frequencies and orientations. We reduce the dimensionality of this large feature output using PCA. In our experimentation, we take the first 10 coefficients as the required features - we have empirically found this value to be suitable for our experiments.

The above procedure establishes the two ends of the semantic axis, described above. Given a new image, we subject it to the same procedure - the first 10 coefficients of a PCA of the output of the Gabor filter bank, and plot the closest point (distance of least approach - the perpendicular from the point to the line in 10-dimensional space). The same image is rated on the above continuity and depth scales by human raters as well.

5 Conclusion

A projected video in a conference room - this paper envisages an immersive experience for participants in the room by 'equalising' geometric and spectral features in the projected video. This paper represents work in progress. We plan to experiment with various histogram features and modelling aspects in the colour space distributions, as inputs to perceptual experiments of visual continuity. We plan to extend our work to multiple displays with geometric correction induced by head movements of participants. Such rendering is possible using binaural cues needed for spatial discrimination [6]. We intend to make the system more robust by planning perception experiments, where the scene is visualised in noisy video environments.

References

1. Criminisi, A., Reid, I., Zisserman, A.: Single View Metrology. International Journal of Computer Vision 40(2), 123–148 (2000)
2. Faugeras, O.: Three-Dimensional Computer Vision. MIT Press, Cambridge (1993)
3. Hoiem, D., Efros, A.A., Hebert, M.: Automatic Photo Pop-up. In: ACM SIGGRAPH (2005)
4. Irani, M., Rousso, B., Peleg, S.: Computing Occluding and Transparent Motions. International Journal of Computer Vision 12(1), 5–16 (1994)
5. Oliva, A., Torralba, A.: Modeling the Shape of the scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
6. Ralph Algazi, V., Duda, R.O.: Headphone-Based Spatial Sound. IEEE Signal Processing Magazine, 33–42 (2011)