14INT009
2014/2/28

解像度を維持しつつ自然な Bullet Time を実現する射影変換の検討

坂本 竜基[1,a)]　陳 鼎[1]

概要：ネットワークにおける情報伝達の手段は、文字から画像、画像から映像へとシフトしつつある。この映像も単一の映像ではなく、多視点映像をもちいた表現も多くなってきた。このうち、被写体を中心に等間隔に並べた多視点カメラからの映像を同時刻のフレームを順に切り替えると Bullet Time と呼ばれるカメラワークをもった映像表現となり、近年よく使用されている。この時、各カメラを被写体の一点が画像の中心となるように設置しなければ不自然な映像となってしまうが、厳密な位置合わせは非常に困難であるため各フレームを射影変換で補正する方法がよくとれれる。しかし、この射影変換は元のフレームを変形するため、そのままでは空白部分ができてしまう。これを回避するには変換後の画像を拡大すればよいが、過度な拡大をすると画像が劣化してしまう。本稿では、この拡大をなるべく抑えつつ、Bullet Time カメラワークとして自然な射影変換の方法を提案する。

Optimized Homography for Natural and High-Quality Bullet-Time Camera Work

Ryuuki Sakamoto[1,a)]　Ding Chen[1]

Abstract: The *"Bullet Time"* camera work is realized with flipping through frames at same moment taken by multi cameras surrounding an object at even distances. For making outcome frames of the camera work, the Homography transformation is adapted for rectifying inaccurate camera poses. Therefore the Homography transformation, however, makes some blank spaces during distorting the frame, the scale up transformation should be applied after that. The scaling up, however, makes the quality of the outcame down. In this paper, we proposed a method to calculate Homography matrices for keeping the quality and naturality of outcome frames of the camera work.

1. はじめに

映画や映像表現の現場で古くから知られる専門知識であるカメラワークは、その選択如何によって受け手の印象に大きく影響を及ぼすことが知られている。このうち、時間を止めて被写体の周りを回っているかのような映像である、俗に Bullet Time と呼ばれるカメラワークは、映画をはじめとしてここ 10 年ほどで映像表現に広く用いられるようになった。Bullet Time は、被写体の周囲に複数台のカメラを等間隔に並べ、ある時刻における各カメラのフレームを端に位置するカメラから順に切り替えることで得られる。ただし、例えば、光軸が交互に上下するといった、各カメラの映像を切り替えてみたときに、それらが滑らかに連続した変化として認知できないようにカメラが配置されている場合は、非常に違和感のある映像となってしまう。よって、各カメラの設置は厳密におこなう必要があるが、実際は、光軸が 3 次元空間中に位置する任意の 1 点を通過するように手動でカメラの向きを調整することは極めて難しい。そこで、ポストプロセスとして各カメラのフレームを、あたかもカメラを正しい向きに厳密に調整して撮影したかのように補正するのが一般的である [1]。

一方で、この補正は、あらかじめカメラを校正した上で適切な射影変換行列を掛けることで実現されるため、元の

[1] ヤフー株式会社
Yahoo Japan Corp. Mid9-7-1 Akasaka, Minato-ku
[a)] ryusakam@yahoo-corp.jp

矩形のフレームは歪んだ四角形に変換されてしまう。そこで、その歪んだ四角形を矩形でクロッピングして、出力したい解像度まで拡大したものを出力とするが、これは結局、元のフレームの構成要素である全画素のサブセットであるため実質的な解像度は低下してしまう。

しかし、過去の研究では、この実質的な解像度の低下を抑制する最適な変換行列に関する研究はなされていない。現在、8~ 12 台程度のカメラを同期させたうえ 30fps 以上で撮影する場合、解像度を XGA 以上にするとシステム全体が高価なものになる。つまり、システムのコストを考えればカメラの解像度は VGA 以下に抑える必要があるため、例えば、いまや FullHD が珍しくない TV や PC に配信することを鑑みれば可能な限り解像度は維持するほうが望ましいであろう。そこで、本稿では、Bullet Time において実質的な解像度の低下をなるべく抑える最適な射影変換を提案する。

2. Bullet Time と射影変換

地面を xyz 座標系における xz 平面（$y = 0$）、天方向を y 軸とした空間に被写体が立っている状態で、N 台のカメラで Bullet Time を実現するカメラの配置を考える。最も単純に実現するには、被写体が内包する 3 次元点 $\mathbf{G}$（以下、注視点と呼ぶ）から等距離かつカメラ間の距離を等間隔になるよう各カメラを並べ、各カメラを $\mathbf{G}$ に向ければよい。つまり、$\mathbf{G}$ を含む xz 平面に平行な平面上に $\mathbf{G}$ を中心とした円を考え、各カメラをピンホールカメラモデルとして捉え、k 番目のカメラの焦点位置を $\mathbf{C}_k$（$k = 1, ..., N$）とした時、$\mathbf{C}_k$ を円周上に等間隔で設置したうえで、$\mathbf{C}_k$ から画像平面の中央を通る直線（以下、光軸とよぶ）が $\mathbf{G}$ を通過するようカメラの向きを変えればよい。

しかし実際は、三脚などを使って設置をする場合、上記の手順に厳密に従うことは不可能である。その主な理由は以下の 3 点である。

イ　実空間における $\mathbf{C}_k$ が判らない
ロ　円周やカメラ間の間隔を厳密に測定できない
ハ　三脚の手動操作では $\mathbf{G}$ を光軸上に設定できない

このうち、イとロに関しては、被写体から比較的遠い位置から撮影するのであれば、カメラの大よその中心と巻尺程度の精度の計測で設置しても問題がない。しかし、ハに関しては、事前にモニタ上に出力した格子線を頼りに時間をかけて調整しても実際は十数ピクセルは誤差が生じてしまう。

しかし、もしカメラが強校正済みであるならば、本来撮影されたフレームを任意の 3 次元点を通過するよう光軸を向けて設置したかのような画像に射影変換する、いわばカメラに仮想的なパン・チルトをさせることができる [2], [3]。よって、厳密に光軸を $\mathbf{G}$ に向ける射影変換を各カメラの出力フレームに適用すればハの問題も解決する。

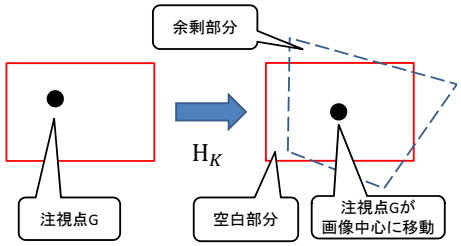

図 1　式 3 による変換
Fig. 1　Conversion by eq.3

この変換をおこなう射影変換行列 $\mathbf{H}_k$（$k = 1, ..., N$）は以下のように求める。まず、各カメラは強校正されているので、k 番目のカメラの内部パラメータ行列 $\mathbf{A}_k$ および外部パラメータのうちの回転 $\mathbf{R}_k$ と並進 $\mathbf{T}_k$ は既知である。このうち、$\mathbf{A}_k$ は以下の要素で構成されているとする。

$$\mathbf{A_k} = \begin{bmatrix} f_k & 0 & u0_k \\ 0 & f_k & v0_k \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

ただし、f_k は焦点距離、$u0_k$、$v0_k$ は画像中心とする。ここで $\mathbf{C}_k = -\mathbf{R}_k^{-1}\mathbf{T}_k$ から $\mathbf{G}$ に向く新しい光軸にあたるベクトル $\mathbf{e_z}$ を求める。$\mathbf{e_z}$ と y 軸との外積から新しいカメラ座標系での x 軸 $\mathbf{e_x}$ が求められ、$\mathbf{e_z} \times \mathbf{e_x}$ から新しいカメラ座標系での z 軸 $\mathbf{e_z}$ を求める。これらから新しい回転行列 $\mathbf{R'}_k$ が以下のように求まるため、

$$\mathbf{R'}_k = \begin{bmatrix} \mathbf{e_x}/|\mathbf{e_x}| \\ \mathbf{e_y}/|\mathbf{e_y}| \\ \mathbf{e_z}/|\mathbf{e_z}| \end{bmatrix} \tag{2}$$

これらから射影変換行列 $\mathbf{H}_k$ が以下のように求まる。

$$\mathbf{H}_k = \mathbf{A}_k\mathbf{R'}_k\mathbf{R}_k^{-1}\mathbf{A}_k^{-1} \tag{3}$$

あとは、これをフレームに適用して変形させれば画像中心に $\mathbf{G}$ が正確に位置する仮想的なパン・チルトした画像が得られる。

しかし、**図 1** にあるとおり変換結果は空白部分を含むため、このままでは出力フレームとすることができない。出力フレームから空白部分を排除する単純な方法は、画像が元のフレームサイズを覆うまで拡大することである。しかしながら、拡大をすれば空白はなくなるが余剰部分が増える、すなわち元の画像にあった余剰部分の画素を捨てることになるため、実質的な解像度は低下してしまう。

この画質の劣化をなるべく抑えるには、元々 $\mathbf{G}$ が映っていた画像上の位置 $\mathbf{g}_k$ に変換後の $\mathbf{G}$ の 2 次元位置が近づくよう画像を平行移動させたうえ、拡大率を余白部分が丁度なくなるようにすればよい（**図 2**）。しかし、各カメラでこれらの最適化を個別におこなうと、隣のカメラへと画像

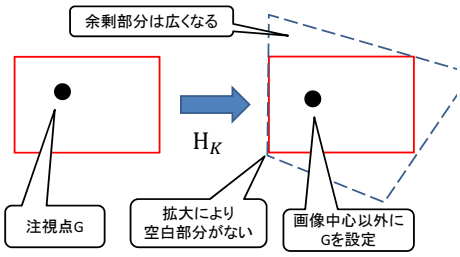

図 2　最適化
Fig. 2　Conversion by eq.3

を切り替えた時に認知的な連続性がなく大変違和感があるカメラワークとなってしまう。本稿では、この違和感をなるべく抑えた上で、実質的な解像度の低下を減らす $\mathbf{H}_k$ を提案する。

3. 拡大と平行移動を含む射影変換

前章で、解像度の低下を抑えるには、拡大と平行移動の二つの手段があることを述べた。アフィン変換における拡大のパラメータは式（1）における f_k、すなわち焦点距離である。内部パラメータにおいて焦点距離が増加するということは、画像面を $\mathbf{C}_k$ から離して $\mathbf{G}$ に近づける、すなわちズームしていることに他ならない。実際には f_k は $|\mathbf{C}_k - \mathbf{G}|$ という距離のみで決定され、$\mathbf{G}$ をその距離に応じた大きさにするのだが、$\mathbf{G}$ は固定であるため裏返せば $\mathbf{C}_k$ を $\mathbf{G}$ に近づけることに相当する。これは、カメラの物理位置を仮想的に前後させることになるため、先行研究では、イ、ロに起因する $\mathbf{G}$ と $\mathbf{C}_k$ 間の距離が各カメラで一致しない誤差を補正するのに用いている [1]。

一方、平行移動は、$(u0_k, v0_k)$ を変更することで達成され、これは $\mathbf{C}_k$ と $\mathbf{R'}_k$ はそのままに画像面を平行移動させることに相当する。すなわち、光軸が画像中心を通過しないカメラとなり、通常にはないカメラの構造になるが、実質的には光軸が画像中心を通った画像と大差がないため認知的な問題は少ない。

これらから、新しい焦点距離 f'_k と新しい画像中心 $(u0'_k, v0'_k)$ が設定された内部パラメータ $\mathbf{A}'_k$ を用いた新しい $\mathbf{H}'_k$ が導出される。

$$\mathbf{A}'_k = \begin{bmatrix} f'_k & 0 & u0'_k \\ 0 & f'_k & v0'_k \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

$$\mathbf{H}'_k = \mathbf{A'}_k\mathbf{R'}_k\mathbf{R}_k^{-1}\mathbf{A}_k^{-1} \tag{5}$$

4. 実質的な解像度の定義

ここで、元のフレームを構成していた全画素が変換後の画像の中にどの程度含まれるのか、すなわち実質的な解像度を計測する指標を作る。画像の四隅の 2 次元座標 $\mathbf{v}_{\alpha k}(\alpha = 0, 1, 2, 3)$ が $\mathbf{H}'_k$ によって $\mathbf{v}'_{\beta k}(\beta = 0, 1, 2, 3)$ へ移動するとする。

$$\lambda \begin{bmatrix} \mathbf{v}'_{\beta k} \\ 1 \end{bmatrix} = \mathbf{H}'_k \begin{bmatrix} \mathbf{v}_{\alpha k} \\ 1 \end{bmatrix} \tag{6}$$

最終的な出力フレームから空白をなくすには、$\mathbf{v}_{\alpha k}$ すべてが、$\mathbf{v}'_{\beta k}$ で形成される四角形の内部に必ず存在する必要があるため、元の画像の全画素のうち、出力フレームに含まれる画素数の割合は $\mathbf{v}'_{\beta k}$ で形成される四角形の面積 S_k と元々の長方形の面積 $S0$ の比 $S0/S_k$ で定義できる。S_k は f'_k、$(u0'_k, v0'_k)$ に依存し、$S0$ は固定値であるため、ここでの問題は

$$S_k = \frac{1}{8}\sum_{\alpha=0}^{3}\sum_{\beta=0}^{3}(\mathbf{v}'_{(\beta \bmod 4)k} - \mathbf{v}_{\alpha k}) \times(\mathbf{v}'_{(\beta+1 \bmod 4)k} - \mathbf{v}_{\alpha k}) \tag{7}$$

という制約条件の下での面積比 E の最大化と定義できる。

$$argmax\ E(f'_k, u0'_k, v0'_k) = \sum_{k=1}^{N}\frac{S0}{S_k} \tag{8}$$

5. 滑らかなカメラワークを実現するアルゴリズム

E を維持しつつ f'_k、$(u0'_k, v0'_k)$ をどう設定すればスムースな Bullet Time になるかを検討した結果、以下の 4 つの戦略を考えた。

戦略 1: $(u0'_k, v0'_k)$ の固定化
$(u0'_k, v0'_k)$ とは、実質的に画像における $\mathbf{G}$ の位置を制御する変数であり、これらを各カメラで共通の値にすれば注視点は見た目上、不動になり違和感がない。この共通の値は、元の画像における $\mathbf{G}$ の位置 $\mathbf{g}_k$ に近いほど E は高くなるため、$\sum_{k=1}^{N}\mathbf{g}_k/N$ と設定できる。

戦略 2: $(u0'_k, v0'_k)$ の変動化
戦略 1 は、$(u0'_k, v0'_k)$ を各カメラ共通とする上では適切であるが、各カメラにおいて $\mathbf{g}_k$ が滑らかに移動しても経験上、違和感はそれほどない。そこで、k が 1 から N まで変化するときの $\mathbf{g}_k$ の軌跡を何らかの線形モデルに回帰させる。直感的に考えると、このモデルは単純であるほど違和感が少ないため実験では直線とした。これは $\mathbf{g}_k$ の分布によっては残差が少なくなり戦略 1 よりも E を増大させる。

戦略 3: フォーカスの固定化
f'_k を増加させれば、擬似的にカメラを近づけたかのような見えになる。よって、f'_k を以下のように設定すれば注視点に居る被写体の見た目の大きさを一定に保つことができる。

表 1　戦略の組み合わせ
Table 1　Combination of each technique

	戦略 3	戦略 4
戦略 1	コンビネーション A	コンビネーション B
戦略 2	コンビネーション C	コンビネーション D

$$
\begin{aligned}
f'_k &= f_{average}\frac{|\mathbf{C}_k - \mathbf{G}|}{z_{average}} \qquad (9)\\
f_{average} &= \sum_{k=1}^{N}\frac{f_k}{N}\\
z_{average} &= \sum_{k=1}^{N}\frac{|\mathbf{C}_k - \mathbf{G}|}{N}
\end{aligned}
$$

被写体の大きさがカメラ間で不規則に大小すると違和感があるため、このような処理は自然な Bullet Time を演出する上では効果的である。

戦略 4: フォーカスの変動化

滑らかに $\mathbf{g}_k$ を変化させた戦略 2 と同じく、被写体の見た目の大きさも滑らかに変化させても違和感は少ない。これは、カメラが被写体を取り囲んでいるような配置を天から見下ろした場合、$\mathbf{C}_k$ を $\mathbf{G}$ と $\mathbf{C}_k$ を結ぶ直線上で前後に動かし、$\mathbf{C}_k$ が全体として何らかの滑らかな軌跡上に存在しているかのように f_k を設定すると達成され、その解の一つは、以下の α を何らかの線形モデルへの回帰させ f'_k を得ることで求まる。

$$
\begin{aligned}
\alpha &= \frac{\mathbf{f_k}}{\mathbf{C}_k - \mathbf{G}} \qquad (10)\\
f'_k &= \alpha(\mathbf{C}_k - \mathbf{G})
\end{aligned}
$$

実験では、天方向から見下ろした 2 次元空間において α を 2 次曲線で回帰させた。この回帰の結果得られる f'_k の残差が戦略 3 で決定した f'_k と f_k の距離の和よりも低い場合は E は戦略 3 より高くなる。

これらの方法を適用した全体の処理は以下のようになる。

(1) ユーザが $\mathbf{G}$ を入力する
(2) 方法 1 か 2 を適用して $(u0'_k, v0'_k)$ を決定する
(3) 方法 3 か 4 を適用して f'_k を決定する
(4) 式 (7) を満たす最小値まで全 f'_k を一定の割合まで増加させる
(5) 各フレームに $\mathbf{H}'_k$ を適用する

このアルゴリズムでは戦略 1、2 および戦略 3、4 はお互いに排他的であるため、実質的に 4 種類の変換行列が出力可能である。以下では、この**表 1** のように 4 種類の戦略の組み合わせをそれぞれコンビネーション A、B、C、D と呼ぶ。

6. 関連研究とマイルストーン

映画ではなく現実に起こった出来事を多視点映像を処理することで効果的に表現するシステムは、古くから応用研究がなされ専らスポーツをその撮影対象に発達してきている [4], [5], [6], [7], [8], [9], [10]。このうち、多視点映像を自由視点映像化する場合は Bullet Time が可能であり、本稿のようにカメラを連続的に切り替える方法は 3 次元モデルを復元する必要がないため最も頑健な手法の 1 つである。

この手法を用いたシステムとして最も著名なのは Eye Vision[11] であろう。Eye Vision は、各カメラの光軸が写したい $\mathbf{G}$ で交差するように、各カメラをロボット雲台で制御することで、アメリカンフットボールの試合中継において Bullet Time を実現した。これに対し、冨山らは、高価で調整に時間がかかるロボット雲台を用いるのではなく、おおよその被写体の位置を向くようカメラを三脚で固定し、位置を合わせをポストプロセスとして画像処理で補正することで仮想的に実現した例を紹介している [1]。この冨山らのシステムは Eye Vision と違い、被写体が大きく動くシーンには適用できないものの、ロボット雲台を導入するよりも圧倒的に設置が容易であるため全日本体操選手権において TV 放送用途での運用が可能であったと報告されている。本稿で提案した手法は、このようなシステムに適用することを想定している。なお、冨山らのシステムでは本稿での戦略 1 で $(u0'_k, v0'_k)$ を画像中心に固定したうえで、戦略 3 を適用するコンビネーション A の特殊ケースとなる射影変換が採用されている。以下の結果や実験では、この冨山らの変換をマイルストーンとして提案アルゴリズムによる E の改善を述べたい。

7. 変換結果

8 台のカメラにより撮影したデータセットに対して、提案アルゴリズムを適用した結果を示す (**図 3**)。図の 1 段目が撮影した元画像であり、画像におけるぬいぐるみの位置や大きさが微妙に違うことがわかる。これに対して、戦略 1、3 の組み合わせであるコンビネーション A で変換した 2 段目を見ると位置も大きさも揃い、Bullet Time の際に自然な切り替えになることが想像できる。3 段目のコンビネーション B は、戦略 3 の代わりに f'_k を回帰により最適化した戦略 4 を用いており、2 段目に比べて大きさが徐々に変化して、全体としてはぬいぐるみが小さく写っている。被写体が小さくみえるということは、実質的な解像度が高いということになり、**表 2** によると E は 0.33 から 0.60 へと改善している。**図 4** は元々の α とコンビネーション A、B の α をプロットした結果であり、コンビネーション B のほうが元々の α に近く滑らかに回帰されていることがわかる。

一方、コンビネーション C は、コンビネーション A の戦

図 3 各シーンの Bullet Time
Fig. 3 Bullet Time camera work on each scene

表 2 各方法によって変換したときの E
Table 2 E values of each technique.

コンビネーション A	コンビネーション B	コンビネーション C	コンビネーション D	冨山の方法
0.33	0.60	0.72	0.78	0.26

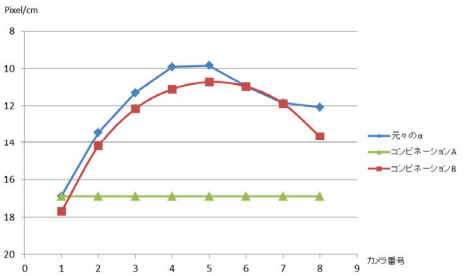

図 4 コンビネーション A、B における α
Fig. 4 α on combination A and B

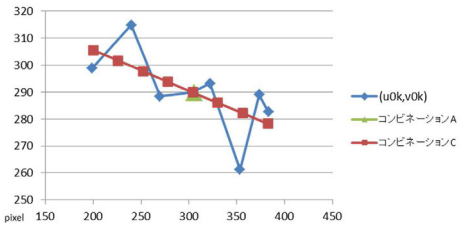

図 5 コンビネーション A、C における $(u0'_k, v0'_k)$
Fig. 5 $(u0'_k, v0'_k)$ on combination A and C

略 1 を戦略 2 に変更し、$(u0'_k, v0'_k)$ を回帰により最適化したものである。図の 4 段目を見ると、この効果により、2 段目と比べて大きさは一定であるが、全体的に小さくなっていることが判り、E は 0.72 へと改善している。**図 5** は、各カメラの $(u0_k, v0_k)$ とコンビネーション A、C による $(u0'_k, v0'_k)$ の位置をプロットしたものであり、コンビネーション C は直線上に回帰していることが判る。最後に 5 段目は f'_k および $(u0'_k, v0'_k)$ 両方を回帰させたコンビネーション D であり、E は全体で最高の 0.78 となっている。これらは何れも既存手法である冨山の方法よりも大きく改善しており、提案手法は元画像が持つ解像度を損なうことなく変換できているといえる。

8. 結論

本稿では EyeVision のように、カメラ映像の切り替えによって Bullet Time を実現するシステムにおける射影変換による補正を解像度維持の面から最適化する 4 種類の戦略を提案した。これらの戦略は特に回転中心が画像中心から離れている場合における解像度の維持に非常に効果的であることを示した。今後は、本提案のような段階的な解法ではなく、これを初期値とした非線形最適化により直接的な解法について研究を進めたい。

参考文献

[1] 冨山仁博，　宮川勲，岩舘祐一：多視点ハイビジョン映像生成システムの試作: 全日本体操選手権での中継番組利用，電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解，Vol. 106, No. 429, pp. 43–48 (2006).
[2] Hartley, R. and Zisserman, A.: *Multiple view geometry in computer vision*, Cambridge Univ Press (2000).
[3] Szeliski, R.: *Computer vision: algorithms and applications*, Springer (2011).
[4] Kanade, T., Rander, P. and Narayanan, P.: Virtualized reality: Constructing virtual worlds from real scenes, *MultiMedia*, Vol. 4, No. 1, pp. 34–47 (1997).
[5] Kitahara, I. and Ohta, Y.: Scalable 3D representation for 3D video in a large-scale space, *Presence: Teleoperators and Virtual Environments*, Vol. 13, No. 2, pp. 164–177 (2004).

[6] Inamoto, N. and Saito, H.: Intermediate view generation of soccer scene from multiple videos, *Proc. on 16th International Conference on Pattern Recognition*, Vol. 2, IEEE, pp. 713–716 (2002).
[7] Hilton, A., Guillemaut, J., Kilner, J., Grau, O. and Thomas, G.: 3d-tv production from conventional cameras for sports broadcast, *Broadcasting, IEEE Transactions on*, Vol. 57, No. 2, pp. 462–476 (2011).
[8] Hashimoto, T., Uematsu, Y. and Saito, H.: Generation of see-through baseball movie from multi-camera views, *IEEE International Workshop on Multimedia Signal Processing*, IEEE, pp. 432–437 (2010).
[9] Kimura, K. and Saito, H.: Video synthesis at tennis player viewpoint from multiple view videos, *Proceedings. VR 2005.*, IEEE, pp. 281–282 (2005).
[10] Tomiyama, K., Miyagawa, I. and Iwadate, Y.: Prototyping of HD Multi-Viewpoint Image Generating System-Live broadcasting use at gymnastics competition (60'th'National Championships)-, *IEIC Technical Report*, Vol. 106, No. 429, pp. 43–48 (2006).
[11] Kanade, T. et al.: EyeVision, *Web, http://www. pvi-inc. com/eyevision*.