第 3 回デジタルコンテンツシンポジウム講演予稿集

HD 映像からの視聴者の注目点を考慮した講義映像の自動生成

桐井　孝嘉† 　篠木　雄大† 　藤吉　弘亘†

† 中部大学工学部情報工学科
487-8501 愛知県春日井市松本町 1200
E-mail: †{kiry,shinoki,hf}@vision.cs.chubu.ac.jp

あらまし　本稿では，高解像度映像上の注目領域をトリミングすることで，仮想的なカメラワークを持つ講義映像を自動生成する手法について提案する．トリミングの際に，講師を捉えるように放送カメラマンの撮影特徴を基にしたカメラワークを用いることで，臨場感のある講義映像の自動生成が可能である．また，講義映像中の視聴者の注目点に基づいた仮想カメラワークを実現することで，視聴者にとって見やすく，学習効率の高い講義映像の自動生成が可能となる．本手法と従来の講師座標のみからなる仮想カメラワーク映像との比較実験を行い，学生 23 人によるアンケート評価の結果，本手法の有効性を確認した．

キーワード　高解像度映像，トリミング，注目点，仮想カメラワーク

Generating a Lecture Video from High Definition Video based on Viewer's Attention

Takayoshi KIRII†, Yudai SHINOKI†, and Hironobu FUJIYOSHI†

† Department of Computer Science, Chubu University
1200 Matsumoto-cho, Kasugai, Aichi, 487-8501 Japan
E-mail: †{kiry,shinoki,hf}@vision.cs.chubu.ac.jp

Abstract　This paper presents a method for generating a dynamic lecture video from the high definition video recorded by a HDV camcorder. Our method generates a lecture video by cropping from the high resolution image to track the region of interest (ROI) such as the instructor. Since the system uses virtual camerawork that is modeled on actual broadcast cameramen's techniques, the generated video is very similar to video shot by a broadcast cameraman. Moreover, by considering the viewer's attention on the chalkboard, the generated video is useful for students. We evaluated our method by subjective experiments, and confirmed the effectiveness of the proposed method.

Keyword　high definition video, cropping, point of gaze, camerawork

1. ま え が き

近年，MIT の OpenCourseWave に代表される，講義とその関連情報のインターネット上での公開活動は，教育効果が期待できるだけでなく，社会貢献の面からも注目されている．特に，講義映像を配信することは，視聴者に講義の雰囲気までも伝えることができるため，その需要は高いと考えられる．また，iPod 等の携帯映像端末へ転送することで，利用者は時間や場所に制約されずに学べる利点がある．

しかし，講義映像を専門のカメラマンや編集者を雇用して作成することはコストの面から難しく，この問題に対して講義の自動撮影に関する研究が国内外で行われているが，既に提案されている自動講義映像撮影システム [2] [3] [4] では専用の設備が必要となる．一方，トリミングによる仮想的なパンニング・ズーミングを実現する手法 [5] [6] が提案されているが，パンニング・ズーミング動作の速度曲線は十分な検討がされておらず，自然なカメラワークとはならない．

そこで，本稿では 1 台の高解像度カメラから臨場感あふれる講義ビデオを自動生成する手法を提案する．ハイビジョンカメラで撮影した講義室前方の高解像度映像から，注目領域のトリミングを行うことで，仮想的なカメラワークを持つ講義映像を生成する．さらに，講義映像中の視聴者の注目点に基づいた仮想カメラワークを実現することで，視聴者にとって見やすく，学習効率の高い講義映像の自動生成が可能となる．提案手法

は，あらかじめ撮影した映像に対して行うオフライン処理であり，生成された映像は Video Podcasting 等を用いてインターネットを通じての配信を行う．

2. 高解像度映像からの講義映像の自動生成

本手法は，ハイビジョンカメラを用いて講義の撮影を行う．対象とする講義形態は，図 1 のように一人の講師が黒板を使用して講義する体面授業であり，プロジェクタのような機材は使用されないものとする．HDV（1080*i*）カメラ一台を講義室の後方部に設置し，黒板全体が入るように画角を調整する．一般的な講義室に設置されている幅 6m の黒板に対して，約 20m の位置にカメラを設置した場合，チョークの文字幅 3mm は画像上で 4 画素以上であり，板書文字の認識が容易である．

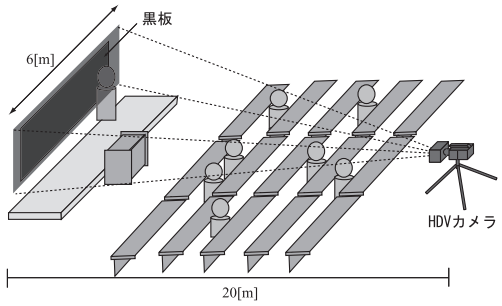

図 1　カメラ配置

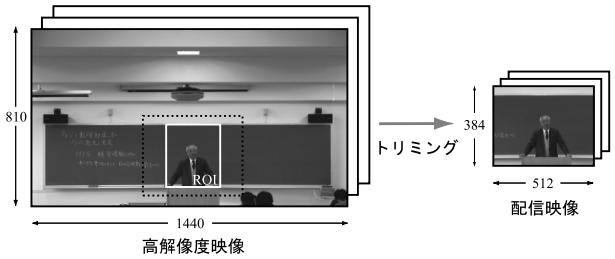

図 2　高解像映像からのトリミング映像生成

撮影した映像を配信する際，ユーザが持つ一般的な携帯映像端末では，HDV カメラで撮影した高解像度映像（1,440 × 810）をそのまま表示することは難しい．そこで，本手法では図 2 に示すように，講師や板書領域等の注目領域に対してトリミングを行うことで，携帯映像端末に適したサイズの講義映像を生成し，配信する．その際，仮想的なカメラワークを決定するトリミング位置やサイズに変化を持たせることで，仮想的なパンニング・ズーミング等の動きを持つ映像を生成することができる．

注目領域座標（ROI）の決定には，講義ビデオ中の視聴者の注目位置（注目点）を考慮することで，視聴者にとって見やすい映像生成を目指す．また，仮想カメラワーク生成の際には放送カメラマンの撮影特徴を基にしたカメラワークモデルを用いることで，臨場感のある映像の自動生成を実現する．

3. 講義ビデオにおける視聴者の注目点

3.1 ポインティングデバイスによる注目点採取

トリミング後の映像に，多数の視聴者の注目点を捉えている講義映像は，多くの視聴者の期待に応えた見やすい講義映像であるといえる．そこで，講義ビデオを視聴するときの視聴者の注目点は，映像上のある 1 点に存在するものと仮定し，ポインティングデバイスを用いて注目点位置の調査を行った．この調査では各 10 分の講義映像を 4 本用意し，20 人の被験者を対象とした．図 3 に (a) 板書時，(b) 板書終了時，(c) 板書消去時，(d) 講師のフレームアウト時の際の視聴者の注目点の分布を示す．図 3 上の丸いマーカで示される点が，それぞれの被験者の注目点を画像上にプロットしたものである．

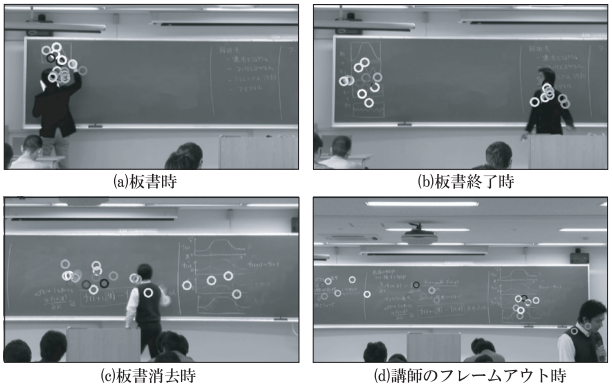

図 3　視聴者の注目点の推移

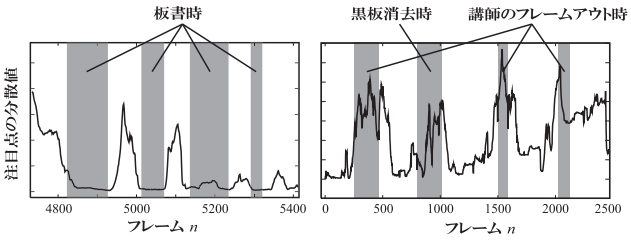

図 4　注目点の分散

図 4 は，講義映像のある 2 つのシーンにおける注目点の分散を示したものである．これらのシーンにて採取した注目点の動きに，以下に示す特徴があることを確認した．

- 視聴者は基本的に講師に注目する
- 講師が板書時，視聴者の注目は板書領域に集中し，板書終了後には板書と講師の 2 領域に集まる
- 板書消し時や講師のフレームアウト時は，視聴者の注目は黒板上に広く分散する

図 4 の分散値は，板書が行われる区間では低く，板書終了直後に高い値を示している．これは，視聴者が新規に書かれた板書を判読し終えるまでの時間に，個人差があるためと考えられる．また，板書消し時や講師フレームアウト時は，それぞれ講義の一区切りである場合が多く，それぞれの被験者の注目点が板書を見直すような動作をしたために，分散が高くなったと考えられる．

3.2 板書への注目固定時間

板書後の一定時間は板書領域に注目が集まるため，板書直後は，トリミング位置を板書領域周辺に固定することが望ましい．そこで，固定時間を決定する注目固定時間モデルを作成する．

採取した注目点データから，板書後に注目が板書から離れるまでの時間を，全ての板書（計 42 回）において計測した．次に，注目固定時間を決める要因として板書の文字量が考えられるため，板書文字ピクセル数と注目時間の関係を調査した．板書文字量は，黒板上の白ピクセル数とし，次章 4.1.2 で述べる講師を消去した黒板画像間の差分から，しきい値処理によって検出したものを用いる．図 5 に注目固定時間と板書文字ピクセル数との関係を示す．

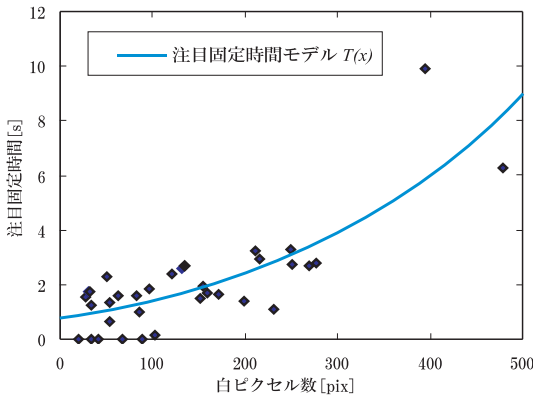

図 5　板書量と注目固定時間

図 5 のプロット点に対して，最小二乗法で 2 次関数をフィッティングした際のパラメータを次式に示す．

$$T(x) = ax^2 + bx + c \qquad (1)$$
$$(a = 2.737 \times 10^{-5}, b = 2.184 \times 10^{-3}, c = 0.834)$$

この時の決定係数は $R^2 = 0.665$ である．(1) 式を注目固定時間モデルとして定義し，板書終了後の仮想カメラワークの決定に用いる．

4. 注目点を考慮した仮想カメラワーク生成

本手法では，高解像度映像からのトリミングを行う際，視聴者の注目特徴を考慮し，かつ放送カメラマンの撮影特徴を基にした仮想カメラワークにより，臨場感のある講義映像を生成する．

4.1 注目領域座標生成

4.1.1 フレーム間差分による講師領域座標の検出

高解像度映像の黒板領域内から講師等の移動体領域を検出する．講義中は板書等により黒板上の状態が常に変化するため，背景推定が対応できない場合がある．

$$\Delta(n) = \max\{\ |I(n) - I(n-k)|\} \quad , 0 < k < 5 \qquad (2)$$

$I(n)$ は現在の入力画像，$I(n-k)$ は k フレーム前の画像とする．急激な輝度値の変化がピクセル上に生じたとき，変化量 $\Delta(n)$ の値は大きくなる．輝度変化をしきい値処理することにより，ピクセルの状態を表す $M(n)$ を求め，移動体を判定する．

$$M(n) = \begin{cases} 1 : object & , \Delta(n) \geqq Th \\ 0 : background & , \Delta(n) < Th \end{cases} \qquad (3)$$

次に，フレーム間差分により移動体と判別したピクセル群をラベリングによりセグメンテーションを施し，各領域を求める．複数の領域が存在する場合，面積が最も大きい領域を講師領域と決定し，その重心座標 $O(n)$=(x_c, y_c) を算出する．

4.1.2 講師消去画像を用いた板書検出

板書検出には，入力画像から黒板上の講師部分を消去した生成画像（講師消去画像）間の差分から板書検出を行う．正確な講師領域を得るため，フレーム間差分の結果を seed とするグラフカット [12] を用いる．

Step1. グラフカットによる講師領域の抽出

フレーム間差分の結果から移動物体と判定したピクセルに対するノードを物体 seed とし，検出領域を膨張した拡散処理により，矩形領域の上，左，右端の領域に対応するノードを背景 seed として設定する．設定した seed を用いて，グラフカットによりセグメンテーションすることで，図 6(d) のような物体マスクパターン $O_{Mask}(n)$ を得る．

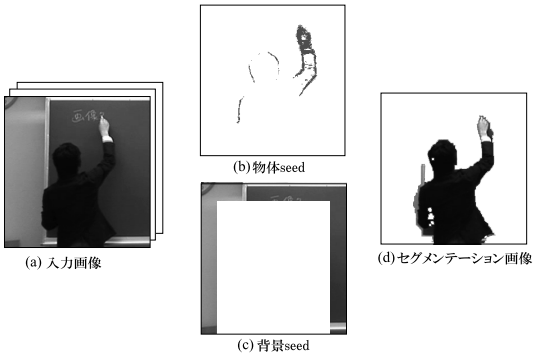

図 6　グラフカットによる講師領域セグメンテーション

$$O_{Mask}(n) = \begin{cases} 1 & : object \\ 0 & : background \end{cases} \qquad (4)$$

Step2. 講師消去画像作成

$O_{Mask}(n)$ により現フレーム画像 $I(n)$ における物体領域以外のピクセルを現ピクセルとして更新することで講師消去画像 $I_C(n)$ を生成する．

$$I_C(n) = I(n) \cdot \overline{O_{Mask}(n)} + I_C(n-1) \cdot O_{Mask}(n) \qquad (5)$$

Step3. 黒板の変化に基づく板書検出

生成した講師消去画像をもとに，新しく板書が書かれたかどうかの判定を行う．以下の式により，黒板上の輝度変化を求め，変化量 $\Delta(i,j)$ をしきい値処理にすることにより板書文字ピクセル $M_C(n)$ を判定し，さらにそのピクセルの総和 $P_C(n)$ を求める．また，その重心座標 $B(n)$=(x_c, y_c) を板書位置とする．

$$\Delta(n) = |\, I_C(n) - I_C(n-1)\,| \tag{6}$$

$$M_C(n) = \begin{cases} 1 : whitepixel & ,\Delta(n) \geqq Th \\ 0 : otherwise & ,\Delta(n) < Th \end{cases} \tag{7}$$

$$P_C(n) = \sum_{(i,j)\in M_C(n)} M_C(n)(i,j) \tag{8}$$

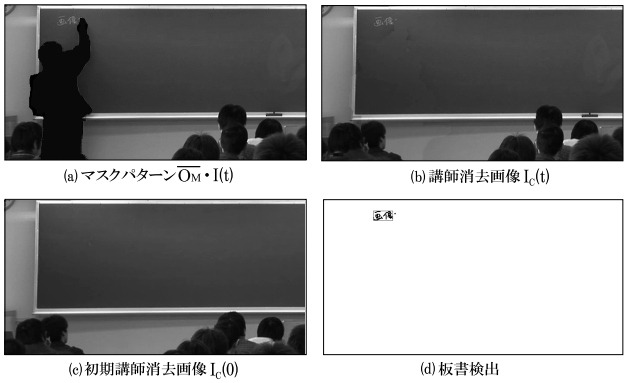

図 7　板書検出例

4.1.3　板書を考慮したカメラワーク生成

仮想カメラワーク算出の基礎となる注目領域座標 $R(n)$ を，講師領域座標 $O(n)$ と 板書位置 $B(n)$，板書終了後における視聴者の注目固定時間モデル $T(P_C(n))$ を用いて次式より算出する．

$$R(n) = \begin{cases} B(n) & : \text{板書検出時} < t < \text{板書検出時} + T(P_C(n)) \\ O(n) & : otherwise \end{cases} \tag{9}$$

なお，$P_C(n)$ は現在時に検出された板書文字ピクセル数である．図 8 に板書が行われたシーンにおける注目領域座標 $R(n)$ の算出例を示す．

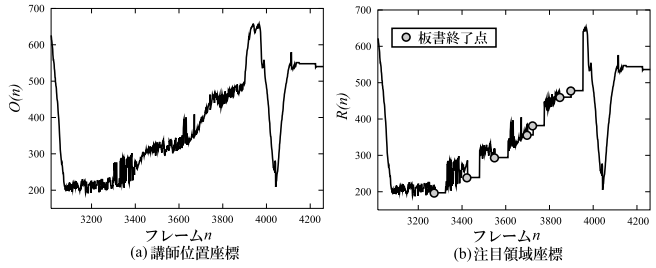

図 8　注目領域座標生成

4.2　仮想カメラワークの実現

4.2.1　バイラテラルフィルタによる変動抑制

注目領域座標のうち，フレーム間差分により得られた講師領域座標部分は，講師の敏速な動きに追従して激しく変動している．このため，本手法ではバイラテラルフィルタ [11] を用いて急激な変化を残しつつ，微細な変動を抑制する．ここでは，次式に示すように一次元に変形したバイラテラルフィルタを用いる．

$$\hat{R}(n) = \frac{\sum_{k=-w}^{w} W(n,k)\cdot R(n+k)}{\sum_{k=-w}^{w} W(n,k)} \tag{10}$$

$$W(n,k) = \exp\{-\tfrac{k^2}{2\sigma_S^2}\}\cdot\exp\{-\tfrac{(R(n)-R(n+k))^2}{2\sigma_R^2}\} \tag{11}$$

4.2.2　カメラワークタイミングの決定

バイラテラルフィルタ適用後の講師領域座標 $\hat{R}(n)$ に対して，零交差処理により特徴点を求め，カメラワークタイミングを決定する．隣り合う各特徴点間の位置座標の変化より，変化の激しい区間をパンニング区間，変化の少ない区間をズームイン区間とする．ズームイン区間を設定する理由は，映像が長期的に変化がない状態を防ぎ，立ち止まって話す講師の発話動作や表情をよく見えるようにするためである．

また，**3.** より，講師が板書を消している時や，フレームアウトしてしまうシーンでは注目点の分散が大きく，黒板の広い範囲に分布していることが確認されている．このため，通常のトリミングサイズ (512×384) では多くの注目点を画面に捉えることが難しい．よって視聴者の注目点に合わせた講義映像を生成するために，これらのシーンはズームアウト区間とし，黒板全体を提示する．

4.2.3　仮想パンニング

テレビ放送では，視聴者が見やすい映像や迫力のある映像を作るために，カメラワークに様々な工夫が施されている．そこで本手法では，臨場感ある講義映像を生成するために，放送カメラマンの撮影技術を考慮した仮想カメラワークを実現する．加藤等は，放送カメラマンのカメラワークを解析した結果，移動体を撮影する際のパンニングに以下の特徴があることを報告している [9].

- パンニング速度曲線は非対称型で，減速時間が加速時間に比べて 6 割程度長い．
- 加速時の最大パンニング加速度は，カメラの動きが最大速度に到達する直前に生じ，減速時には，最大速度から減速に入った直後に見られる．

上記に示した放送カメラマンの撮影技術を考慮した仮想カメラパンニングを実現する．その手順を以下に示す．

Step1. パンニング区間の開始フレームを n_s と終了フレームを n_e としたとき，減速時間が加速時間に比べて 6 割程度長い特徴に従い，加速終了点 n_t を決定し，その座標値 $\hat{R}(n_t)$ を求める．

$$n_t = \frac{0.6n_s + 0.4n_e}{0.4+0.6} \tag{12}$$

$$\hat{R}(n_t) = \frac{0.6\hat{R}(n_s) + 0.4\hat{R}(n_e)}{0.4+0.6} \tag{13}$$

Step2. パンニングの加速度 α は加速区間 $(n_s < n \leqq n_t)$ と減速区間 $(n_t < n \leqq n_e)$ に応じて以下の式により計算する．

$$\alpha = \begin{cases} \dfrac{2\cdot(\hat{R}(n_t) - \hat{R}(n_s))}{(n_t - n_s)^2} & ,n_s < n \leqq n_t \\ \dfrac{2\cdot(\hat{R}(n_e) - \hat{R}(n_t))}{(n_e - n_t)^2} & ,n_t < n \leqq n_e \end{cases} \tag{14}$$

Step3. 高解像度映像からのトリミング位置 $R'(n)$ を，step2 で求めた加速度から計算し，仮想カメラパンニングを実現する．図 9（下）にパンニング速度の変化を示す．

$$R'(n) = \begin{cases} \frac{1}{2}\alpha\cdot(n-n_s)^2+\hat{R}(n_s) & ,n_s<n\leqq n_t \\ \alpha\cdot(n_t-n_s)^2+\frac{1}{2}\alpha\cdot(n_e-n_t)^2+\hat{R}(n_t) & ,n_t<n\leqq n_e \end{cases} \quad (15)$$

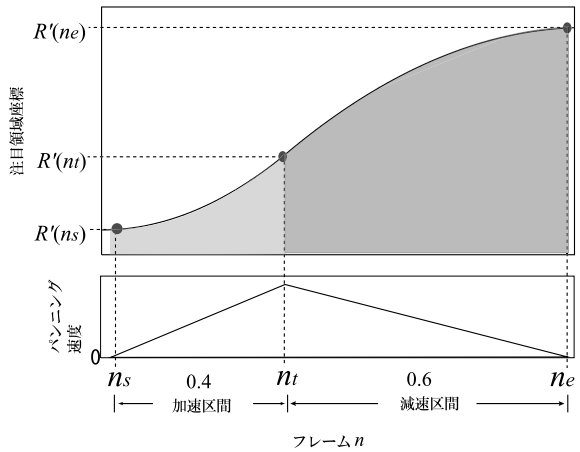

図 9　仮想パンニングモデル

Yokoi 等の講師位置のみからなる仮想パンニング [1] に対して，本手法では板書位置を考慮しているため，板書区間では得られる動作特徴点に違いが生じ，比較的ゆっくりとしたカメラの動きとなる（図 10）．これにより，視聴者は余裕を持って板書を読むことができ，講義ビデオとしての見易さの向上が期待できる．

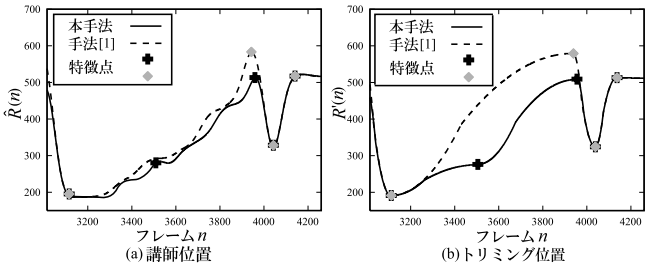

図 10　トリミング位置生成

パンニング区間が長期間に亘る場合，仮想パンニングモデルからなるトリミング位置 $R'(n)$ と講師位置 $\hat{R}(n)$ との相違が大きくなり，画面から講師が外れてしまう場合がある．このような場合は，以下の式によりずれ量 $d(n)$ を求め，画面から外れかけていると判定されれば，ずれ量を基に補正されたトリミング位置 $R''(n)$ を求める．講師の逸脱を判定するしきい値 Th は，トリミング枠幅と，左右に設ける 50〜100 ピクセル程度の安全領域により決定しておく．

$$d(n) = \hat{R}(n) - R'(n) \quad (16)$$

$$R''(n) = \begin{cases} R'(n) & : |d(n)| < Th \\ R'(n)+d(n) & : |d(n)| > Th \end{cases} \quad (17)$$

これにより，講師の見失いを完全に防ぐことができる．ただし，3. に述べたように，板書中，または板書が行われた直後には，視聴者の注目は講師自身よりも書かれた文字に集まることが確認されているため，このような補正は行わない．

4.2.4　仮想ズーミング

放送カメラマンのズーミングにおけるカメラワークの特徴として，石川等はズーミングの加速時間・被写体のサイズ変化に以下の傾向があることを報告している [10].

- ズーミングの加速時間と減速時間の平均値は加速時間がズーミング時間の 47〜60%に対し，減速時間は 40〜53%である．ズームアウトの場合，加速時間はズーミング時間の 38〜63%で減速時間は 37〜61%の範囲である．
- 被写体サイズの変化率の最大値は，ズームイン時には後半に，ズームアウト時には前半に発生するようにズーミングを行っている．

これらの特徴を考慮したズーミングを実現するために，トリミング枠の拡大率を次式より決定する．ズームインもしくはズームアウトにかかる時間を t_z とし，ズーム最大拡大率を z_{max} とする．

$$Z(t) = \begin{cases} 1+\sum_{i=-t_z}^{t} \frac{i^2}{M}\exp(\frac{-i^2}{N}) & ,-t_z<t\leqq 0(zoomin) \\ 1+(z_{max}-\sum_{i=0}^{t} \frac{i^2}{M}\exp(\frac{-i^2}{N})) & ,0<t\leqq t_z(zoomout) \end{cases} \quad (18)$$

M は主にズームイン時の拡大率に影響し，N はズーム時間に影響する．今回は，ズーミングは約 1[sec] の間に黒板の上下が画面に十分映る大きさまでおこなうため，$M = 1800, N = 130$ の値を用いる．図 11 に提案する仮想カメラワークにより生成された講義映像のパンニング，ズームイン，ズームアウト例を示す．

5. 評 価 実 験

5.1　注目点による評価

視聴者の注目点を考慮して生成された仮想カメラワークが，実際にどの程度の注目点を捉えているかについて評価する．あらかじめ 20 人分の注目点データを採取した，8 つの板書を書き終えるシーンに対して，講義映像を自動生成した結果，生成された画像中心と注目点座標との距離をヒストグラム化したものが図 12 である．これにより，トリミング後の映像がどの程度の被験者の注目点を捕らえているかを知ることができる．比較を行う対象として，講師領域座標のみからなる従来の Yokoi 等の仮想カメラワーク [1] を用いた．図 12 より，提案法による映像は平均距離および分散値が従来法に比べ小さく，視聴者の注目点がより画像中心に近い位置に分布していることがわかる．これにより，講師領域座標のみからなる従来のカメラワークよりも，視聴者の注目点に近い映像を生成できた．

5.2　アンケートによる評価

主観評価として，アンケートによる評価を行った．同一シーンにて，本手法により生成された講義映像と，講師領域座標のみからなる従来の講義映像を 10 シーン用意しておき，順不同

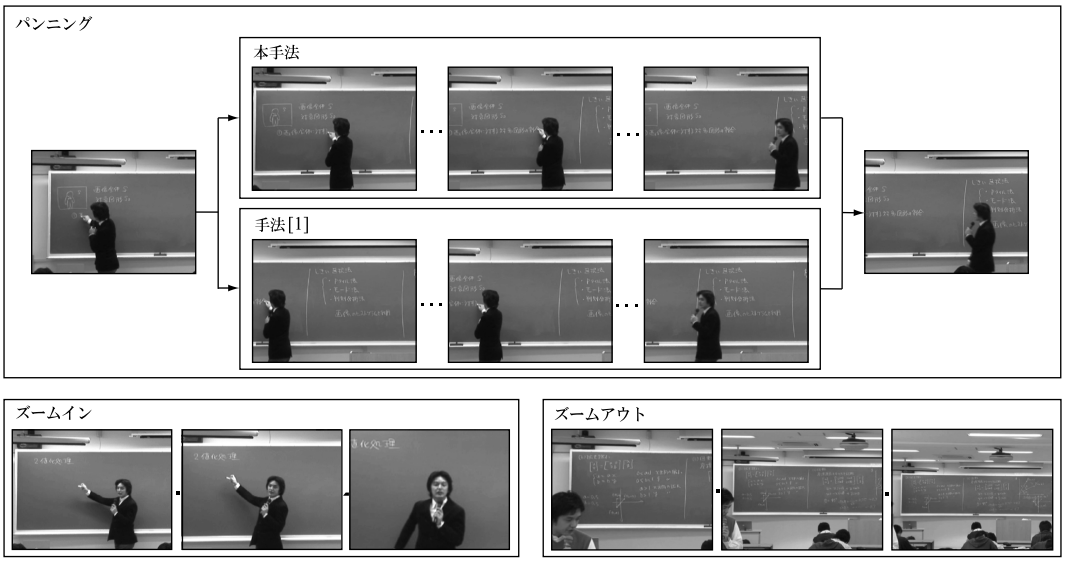

図 11　仮想カメラワークによる講義映像生成

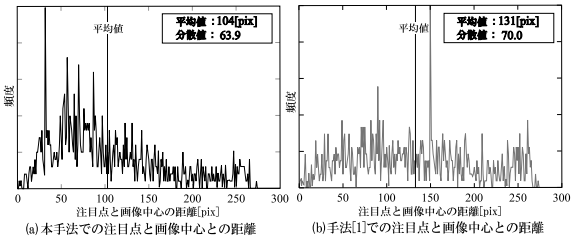

図 12　注目点と画像中心との距離ヒストグラム

で 23 人の被験者に提示した．被験者には，見易さの観点からどちらの映像が優れているかを選択してもらい，違いがわからない・どちらも良くないといったような意見に対しては，”どちらでもない ”という項目を用意することで対応した．

表 1　アンケートによる得票割合

	提案法	従来法 [1]	どちらでもない
得票率 [%]	56.1	32.1	11.7

アンケート調査から，過半数の 56.1% の視聴者から従来法よりも良いという評価を得ることができた．また，t 検定（有意水準 5%）の結果，両者の間に有意差があることが確認されたため，本手法による講義ビデオは従来法よりも見やすい講義ビデオの自動生成をすることができたといえる．

6. むすび

本稿では，視聴者の注目点を考慮した講義映像生成法について提案した．講義映像中の視聴者の注目点を調査した結果，得られた視聴者の注目特徴を基にトリミングを行うことで，視聴者にとって見やすい映像生成が可能となり，放送カメラマンの撮影手法に基づく仮想カメラワークを用いることで，臨場感ある滑らかなカメラワーク映像を実現することができた．アンケートによる評価実験から従来法との比較を行い，講義映像の見やすさに対して，その有効性を確認した．

今後の課題としては，ユーザの入力に対してのカメラワーク変更などを可能とするインタラクティブ機能を含めたサーバクライアントシステムの構築や，パワーポイント等の他の講義形態への対応の検討が挙げられる．また，自動検出する講義イベントとして，図を指し示す等の講師の姿勢を考慮することで，より多彩な仮想カメラワークを実現する予定である．

文　献

[1] T. Yokoi and H. Fujiyoshi, “Virtual Camerawork for Generating Lecture Video from High Resolution Images”, Proc. of IEEE ICME 2005,(July.2005).

[2] 西口 敏司, 亀田 能成, 角所 考, 美濃 導彦 “大学における実運用のための講義自動アーカイブシステムの開発” 信学論 (D-II), Vol.J88, No.3, pp.530-540(2005).

[3] 尾関 基行，中村 裕一，大田 友一，“机上作業シーンの自動撮影のためのカメラワーク”, 信学論 (D-II), Vol.J86，No.11, pp.1606-1617(Nov.2003).

[4] 尾関 基行，中村 裕一，大田 友一，“注目喚起行動に基づいた机上作業映像の編集”, 信学論 (D-II), Vol.J88，No.5，pp.844-853(Nov.2005).

[5] 大西 正輝，泉 正夫，福永 邦雄，“ディジタルカメラワークを用いた自動映像生成”, 画像の認識・理解シンポジウム (MIRU’2000) 講演論文集 I, pp. 331-336(2000).

[6] 大西 正輝，村上 昌史，福永 邦雄，“状況理解と映像評価に基づく講義の知的自動撮影”, 信学論 (D-II), Vol.J85，No.4, pp.594-603(Nov.2002).

[7] A. Lipton, H. Fujiyoshi, and R. Patil, “Moving target classification and tracking from real-time video”, Proc. of WACV, IEEE, pp.8-14(1998).

[8] 西口敏司，仙田修司，美濃導彦，池田克夫, “首振りカメラによる黒板の記録手法”, 画像の認識・理解シンポジウム (MIRU’96) 講演論文集 I, pp.37-42(1996).

[9] 加藤大一郎，山田光穂，阿部一雄，“スタジオ番組における放送カメラマンのカメラワークと視線の動きの分析”, テレビ誌, Vol.49, No.8, pp.1023-1031(1995).

[10] 石川秋男, 加藤大一郎, 津田貴生, 福島宏, 下田茂, 阿部一雄, “放送カメラマンのズーミング計測法の検討と静止している被写体を撮影するときのズーミング解析”, 映情学誌, Vol.53, No.5, pp.749-757(1999).

[11] C. Tomasi and R. Manduchi, “Bilateral Filtering for Gray and Color Images”, Proc. of ICCV, pp. 839-846(1998).

[12] Y. Boykov, M. Jolly, “Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images”, Proc. of ICCV, vol. I, pp. 105-112(2001).