社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－CE－81 (6)
2005／10／22

複数の自動追尾カメラを用いた映像の安定度と連続度に基づく映像切り替え方式の提案

宮下 剛† 品川 高廣† 吉澤 康文†
† 東京農工大学

e-Learningに利用される講師映像は単一の固定カメラにより撮影した映像が多い．このような映像は単一アングルによる単調な映像の連続となり，受講者は飽きを感じてしまう．筆者らは講師を自動的に追尾する複数のカメラを用いて，映像の安定度と連続度に基づき映像編修する方式を提案する．本稿では，この提案方式により得られた映像をネットワーク経由で配信するシステム:sCam の設計，実装，評価について述べる．

An Editing Strategy for e-Learning Based on Consecutive Stability of Images with Automatic Shooting Cams

Tsuyoshi Miyashita †, Takahiro Shinagawa † and Yasufumi Yoshizawa †
† Tokyo University of Agriculture and Technology

Movies used for e-Learning are mainly shot by a single fixed camera and from a single angle, thus becoming monotonous for those who attend the classes. We have proposed Smart e-Learning Camera System (sCam) for e-Learning, with several cameras for automatic shooting of the lecturer and the editing strategy for consecutive stability. In the system, video sources are collected via TCP/IP network. We here verified effectiveness of our proposal.

1. はじめに

現在，e-Learningの導入，運用が進むにつれ，省力化・低コスト化が望まれている[1]．例えば，講義の様子をわかりやすく撮影するためには，講師や黒板などを的確に撮影するカメラマンや，複数のカメラ映像から適切な映像を選択するスイッチャーなどが必要となる．しかし，大学内のいたる所で数多く行なわれる講義を撮影し，e-Learning用に編集を施して配信することは，人手，時間，コスト面で負担となる．そこで，撮影を省力化するため，プロカメラマンのような撮影知識を計算機に組み込み，自動的に最適な映像を取得する手法が提案されている．

これらの手法[2,3,4]は，大きな教室で大人数の学生を対象にして行われるマス形式の講義の撮影を目的としたものが多い．このため，研究室で開かれるセミナー形式の講義のように，小さな教室で少ない人数を対象として行われる講義の自動撮影には向いていない．例えば，従来の手法ではカメラ位置があらかじめ固定されていることを前提としていため，小さな教室で必要に応じてカメラを設置するといった利用法が難しい．すべての教室にあらかじめ複数台のカメラを常時固定設置しておくことは，コストの点からも現実的ではない．また，セミナー形式の講義では，固定された黒板だけでなく，必要に応じて可動式のホワイトボードやPCによるプレゼンテーションソフトを組み合わせて利用することがあり，背景が固定した映像を前提とした手法でイベントを抽出することは難しい．さらに，セミナー講義は大学教育で重要な位置をしめており，e-Learningにおいてもセミナーを取り入れる必要性は大きい．

しかし，実際の講義映像の撮影は，講師や学生により，高々1台程度のDVカメラで行われ，編集はタイトルカットやロゴを挿入するといった程度のものが相場である．その講

師や学生は撮影，カメラワークに関する特別な教育や訓練は受けていない．このため，カメラワークなどにおいて質の高いものは望めない．

このように撮影された映像は，講義を行っている講師の顔，振舞い，所作が主なものであるが，受講者がこれらの映像（以下，講師映像と呼ぶ）を見た場合，飽き易く，単調なものとして感じられることが多い．

この主たる原因として e-Learning 用に撮影される映像が，講師の上半身をフレームの中央に配置した単一視点，単一アングルによるものがほとんどだからである．これに対し，映画や TV 番組の映像は，ディレクタによる，何をどのように撮影するかという具体的な指示であるシナリオ（カメラ割り）をもとに，カメラワークの知識を持つカメラマンにより複数のアングルから撮影される．さらに，撮影された映像は，ショット間の意味，時間，リズム感を考慮しながら切り替えるカット編集を経ることで，視聴者にとって飽きにくい単一の映像に仕上げられている．

そこで筆者らは，撮影対象として小教室にてセミナー講義を行う講師を取り上げ，ユーザが自由に持ち運び可能な 2 台のアクティブカメラで講義を自動撮影し，2 つの映像の中から最も適した映像をリアルタイムに選択して配信するシステム（Smart e-Learning Camera System : sCam）を提案する．提案手法では，背景に関する予備知識を必要としない画像処理手法により講師を追跡しながら撮影するとともに，カメラの動作状況や映像編集の経験則などに基づいて適切な映像をリアルタイムに選択して，講義に適した映像を配信する．

本稿では，まず 2.で関連研究と映像編集における基礎知識を述べ，3.で試作システムの全体像と，そのコンポーネント技術については 4.で，最後の 5.では試作システムによる撮影実験の結果と考察についてまとめる．

なお，本稿では，映像の最小単位をショットと呼ぶ．ショットは，ある 1 つのカメラの撮影による映像単位とし，そのショット内には不連続な映像は存在しないものとする．また，カットは，単一で使用している場合，ショットと同義とする．

2. 関連研究および既存技術

2.1 複数視点による講師の自動撮影

撮影対象となる講義数の増加により，講師映像生成の省力化の要求が高まってきており，カメラマンが撮影するような質の高い映像を自動的に生成する研究が行なわれている．また，複数のカメラを用いて撮影を行なう場合でも，受講者が同時に見ることのできる映像は 1 つであるため，複数のカメラから得られた映像の中から，最適な映像を選択する手法も提案されている．これまで提案されている複数のカメラを用いた自動撮影の研究 [2]では，講師の位置情報，その講師を撮影する具体的な指示が前提知識として，トップダウンで与えられている．そのため，撮影が行なわれる環境や状況が変化した場合，撮影ルールの再構築が必要となり，汎用的な自動化は難しい．

そこで，講師の動作や所作などを事前に，運動履歴として収集し，この運動履歴から，設定した映像評価基準を満たす固定ショット撮影ルールを用いて複数カメラの固定ショットを切り替えることより撮影を行なう手法も提案されている [3]．

これらの研究は，大学の比較的規模が大きい教室や講堂などで行なわれる撮影を想定し，撮影する対象も講師だけでなく，学生や板書内容も含まれている．様々な情報をまんべんなく獲得するために，教室中に撮影用カメラ，観測用カメラや，魚眼カメラを数多く配置する必要がある．また，講師の位置抽出や，運動履歴の獲得のために，位置センサを講師に装着しなければならない．

これに比べ，大西らによる研究 [4]は 3 台のカメラで撮影する．教室の黒板から 6m 程度

後方で，左右・中央に固定カメラを設置し，それぞれのカメラによって取得された映像から，講師映像として最も適しているショットの出力カメラをスイッチングすることにより選択している．基本的なポリシとして，黒板を利用した講義での講師映像の撮影としているため，講師は背景に黒板がフレームされない場所での自動撮影はできない．

以上を踏まえ，sCam は自由に動き回る講師 1 人の自動撮影をとりあげる．ユーザにより自由に持ち運び可能なアクティブカメラで広い範囲で追跡撮影をする．このとき，背景に関する予備知識を必要としない，汎用的な自動撮影方法の実現を目指している．

2.2 映像編集

一般に講師映像は，同じアングルにより撮影したものを，利用されることがほとんどである．この理由は，複数のカメラを所有していないことや，教員自らや学生によって，講義が始まる直前に，カメラを教室に持ち込み，出席している学生の邪魔にならないように最後尾から，なるべく広範囲をフレームに入れるようにアングルを決定するこが原因と考えられる．従って，撮影された映像は単一視点となり，時間短縮以外に映像編集の余地がない．

一方で，映画やテレビ番組の映像は，複数のカメラによって撮影された映像の中から，適切なショットを選択し，切り替えながらつなぐことによりカット編集されている [5]．例として講義の場合は，講師の映像，教室全体の映像，質問者の映像，スライド映像，板書映像などを，時間やショット間で矛盾がないように，切り替えながらつなげることが基本とされている．

しかし，1 台のカメラによる撮影では，複数の視点映像を獲得することができず，敢えて編集をしようにも，一般のビデオ編集で行われているような，同じサイズで単調な映像の連続にならざるをえないのが現状である．

sCam の狙いは，さまざまな撮影位置からの映像を切り替えながらカット編集ができるようにすることにある．筆者らは 2 台以上のカメラを用意することにより，異なる視点，構図，画角を有する映像を一気に撮影し，そのなかから適切と判断する 1 映像を適宜選択し，それらをつなぐことでリズミカルな映像に仕上げることができる可能性が高いとした．

3. プロトタイプシステム

3.1 システムの構成

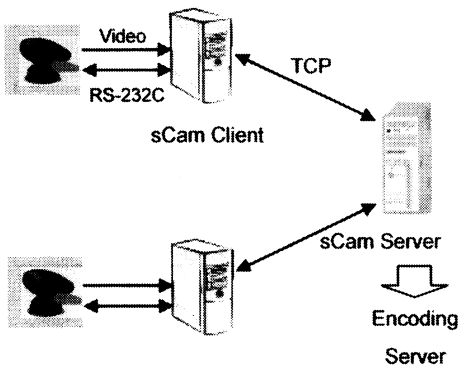


図 1　システムの構成

sCam のシステム構成を図 1 に示す．sCam は講師の撮影を行なうカメラ 2 台，このカメラを接続したクライアント PC (以下，sCam クライアントと呼ぶ) 2 台，サーバ (以下，sCam サーバと呼ぶ) 1 台，そしてネットワークから構成されている．

表 1　カメラの役割と配置

カメラ	役割と配置
Master カメラ	講師を含めた講義全体の様子を撮影する．Sub カメラよりも，後方に配置する．
Sub カメラ	主に講師の追跡撮影に使用する．講師により近いほうに配置する．

カメラは講師の追尾撮影のために，パン(左右)・チルト（上下）などのカメラの向きや，ズームの倍率を PC で制御できるアクティブカメラ（Canon コミュニケーションカメラ

VC-C4）を使用した．このカメラの配置を基本的にユーザが自由に決めることができる．ただし，sCam では 2 台のカメラそれぞれに表 1 のような役割を持たせており，その配置も若干の制約がある．実際に配置した例を図 2 に示す．

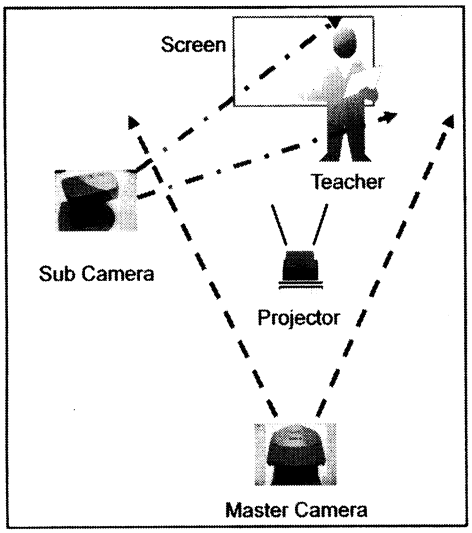


図 2　カメラの配置例

また，3 台以上のカメラを用いることも可能であるが，カメラ台数を増やすこととは，学生の座る場所や発表場所などで不便になると予想したため，本試作ではカメラの台数を 2 台とした．

sCam クライアントは，PC とアクティブカメラの 2 つで構成する．クライアントは，カメラからアナログビデオ出力をビデオキャプチャボードで入力，さらに RS-232C 経由でパン・チルトなどの制御コマンドを送ってカメラのアングルを制御する．

sCam サーバは，Master カメラと Sub カメラの 2 台のカメラ間の協調を行なうために，sCam クライアントは，画像処理の結果やカメラの動作状況などの情報を送信する．そして，サーバは 2 台のクライアントから受け取った情報をもとに，適切な映像を決定して，その切り替えを行なう．

クライアント・サーバ間は 10/100base スイッチングハブを介して Ether ケーブルで接続する．

なお，クライアントとサーバのいずれも，OS に RedHat9（Linux Kernel2.4.20-8）をインストールした PC（CPU：Intel Pentium4 2.4GHz ，Memory：1GB）を使用した．

3.2 自動講師追尾

クライアントに接続されたカメラは，それぞれ撮影した映像をリアルタイムに画像処理することにより，自動的に講師を追跡する．講師が移動した場合には，その移動方向や移動量を認識してパン・チルト制御を行ない，フレーム内に講師が入るようにカメラの向きを調整する．また，講師がフレーム内に適切な大きさで写るように，自動的にズーム制御も行なう．

ただし，追跡撮影を実現する画像処理やカメラワーク制御などを行なうために，表 1 で示したように，2 台のカメラにはそれぞれの固定した役割が与えたれている．Sub カメラは講師近くに配置され，フレームの中央付近で，適正なサイズで捉えるようにパン・チルト・ズーム制御をこまめに行なっている．そのため，撮影される映像はスピード感やダイナミックな効果を得られるが，反面ブレ感や焦点ボケが生じやすいという欠点がある．対照的に，Master カメラは，講師を含めた周辺環境をなるべく広くフレームに収めるために，画角を広くしている．

また，Master カメラも講師がフレームを外れそうになったときにパン・チルト制御を行なっているが，このときの制御量は Sub カメラの制御量よりも小さくすることにより，映像が急激に変化することを抑えている．Master カメラからの映像は「ひき」気味の映像となるため，視聴者に対して客観的で冷静な印象を与える特徴を有する反面，映像に動的な要素が小さいため飽き易く，面白みがないという欠点がある．

3.3 講師追尾のカメラワーク方法

前節で画像処理により認識された講師領域が映像の中心に収まるようにするために，講師の重心座標が中心付近に近づくように定期的にカメラのパン・チルト制御を行う．あまり頻繁に制御を行うと，カメラが小刻みに動いて映像が見にくくなるので，講師領域の重心が映像の中心からある程度以上離れたときにのみパン・チルト制御を行う．

講師領域を中心付近に捉えた後は，話者が映像内で適切な大きさになるようにズーム動作を行う．カメラの 16 倍率ズームを 16 ステップに分割し，フレーム毎のズーム倍率を 1 ステップ，1 倍率までと制限することにより，頻繁な制御を抑え，急激なズームイン・ズームアウトを避けるようにしている．

3.4 カメラ間の協調

Master カメラと Sub カメラを協調させることにより，状況に応じて最も適した映像を選択出来るようにする．例えば，3.2 で述べた Sub カメラによる自動講師追跡は，背景がブレることにより落ち着きのない映像になる．そもそも，パン・チルトなどのカメラワークは，そこに明確な意図（スピード，臨場感，没入感等）があることを前提としており，不安定な映像とのトレードオフを常にカメラマンは意識しなければならい．したがって，明確な意図がない場合には，固定ショットによる撮影が基本である．

また，講義・講師映像などは，その内容にも左右されるが，一般的に退屈になりがちで，視聴者の注意や興味をつなぎ止めることは難しい．たとえ演劇や舞台等の撮影であっても，固定アングルから同じ映像が長時間続くと観客が飽きてくることは，映画の進化過程で明らかになっている．そこで，視聴者への訴求力を高めるため，ショットを切り替えたほうがよいことが経験則として知られている [6]．このような条件を考慮して，2 つに映像を切り替えることにより，カット編集を実現し，同時にカメラ制御によって生じる不安定な映像の低減を行ない，視聴者に呈示する映像をより自然なものにする．

4. 映像切り替え

sCam では，Master カメラと Sub カメラの映像を適切に切り替えることによりカット編集を生み出している．同時にパン・チルト動作中の映像を見せないようにする目的もある．このため sCam は，カメラのパン・チルト制御量，ズームの有無，ショット時間の 3 つを利用する．

4.1 パン･チルト制御量

sCam は動いているカメラの映像を見せないようにするため，パン・チルト制御量を用いる．制御量の値が小さいほどカメラの動きが小さいため，より見やすい映像になると考えられる．

制御量の算出は，画像処理により講師と推定する差分領域の重心座標と，sCam で非追跡領域とよぶ赤枠の内部で x 方向，y 方向のそれぞれを任意に設定された座標までの差分画素の合計である．

たとえば，カメラが時刻 t において，図 3 の左のような位置で講師を認識した場合，そのモーメントを G(x, y)とする．これを時刻 t+⊿において G'(x, y)と重なるように，カメラの視点をパン・チルト制御することにより追跡している．言い替えれば，講師の重心座標が非追跡領域を，x 方向と y 方向に外れた分（図 3 では双方向の矢印で結ばれた画素分）を元に戻すフィードバック制御である．

この制御量の値は Sub カメラと Master カメラのそれぞれで，講師が非追跡領域を外れるたびに計算され，その結果をサーバへ送信する．サーバは 2 つのクライアントから送られてきた制御量の値を比較し，より値が小さいほうのカメラ出力映像を配信する映像として選択する．

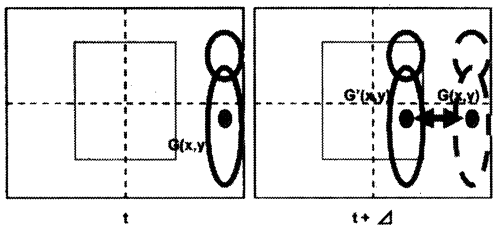


図 3　パン・チルト

4.2 ズームの有無

sCam ではズーム動作によって得られる映像を極力排除している．オートフォーカスを利用しつつ，自然なズームを実現することは困難である．また誤認識による，意図しない部位へのズームが発生するなど問題が多い．このため，ズームイン・アウトの過程は視聴者にズームイン・アウト後にオートフォーカスが働くまでの時間を制御することができない．このため視聴者にとって見にくい映像となる．そこで，ズーム処理が行なわれた場合には，強制的に，ズームを行なわないカメラ出力映像へ切り替えるようにする．

4.3 ショット時間

制御量とズーム判定を組み合わせでは，講義の全景を撮影する Master カメラよりも，講師の近くに配置されている Sub カメラの方が制御量の値は大きくなりがちである．その結果，Master カメラばかりが選択されて，退屈な映像になってしまう恐れがある．そこで，常に一方のカメラが選択されつづけることを防ぐため，2 台のカメラ出力であるショットに，表 2 のような最小と最大のショット時間を割り当てている．

表 2　ショット時間

	Master カメラ	Sub カメラ
最小時間	5s	8s
最大時間	10s	16s

これらの時間の根拠は，映像編集における経験則に基づくものである．あるシーンを説明するようなロングショットやマスターショットとよばれるものは，8〜10 秒程度とされ，顔から胸あたりをフレームするようなミドルショットは 5〜6 秒程度とされている．sCam では，Master カメラによるショットが，ロングショットに，Sub カメラのショットが，ミドルショットに相当する．最大時間は最小時間の 2 倍とした．

4.4 切り替え手順

講師が動いている場合に，映像切り替えがどのように行なわれるか具体的に例を示す．

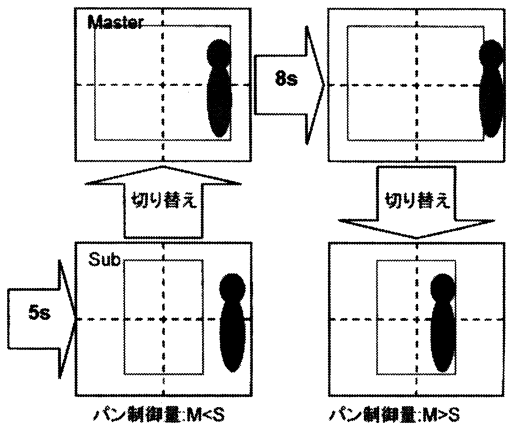


図 4　講師移動時の切り替え

まず，図 4 では，上側が Master カメラの映像，下側が Sub カメラの映像を表している．図 4 の左フレームの時点で話者を認識したとき，Sub カメラの制御量の値は Master カメラのそれよりも大きくなる．そこで，Master カメラから撮影されたショットを視聴者に呈示する映像として選択する．一旦，Master カメラが選択されると，4.3 節のショット時間に従い，制御量の値にかかわらず最低 8 秒間はそのカットを選択し続け．その後，Sub カメラの制御量の値が小さくなったら，再び Sub カメラを選択する．

講師が静止している場合は，いずれか一方のカメラを選択し続けるのを防ぐために，最大のショット時間を過ぎると強制的にカメラを切り替える．例えば図 5 では，Master カメラを 16 秒間選択し続けた場合は，制御量

の値にかかわらず Sub カメラの映像に切り替える.

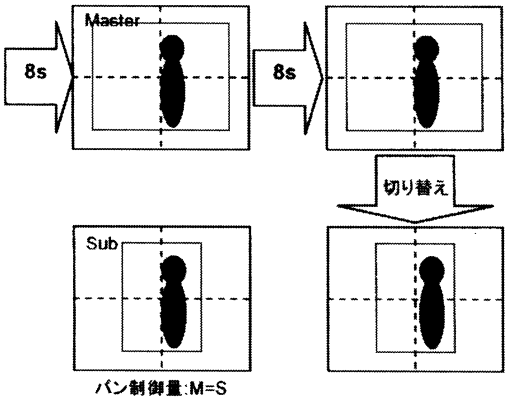


図 5 講師静止時の切り替え

また, Master カメラ, Sub カメラともに, 最小時間を越えた場合は, 最大時間まで, 1 秒ごとに制御量の比較を行ないながら, カメラを切り替える.

5. 実験と考察

sCam による, Master カメラから Sub カメラへ, そして Sub カメラから Master カメラへの切り替えが発生した時のショット時間ごとの頻度を調べる. これは, 講師映像において, 切り替え時におけるショット時間の傾向を確認することが目的である.

5.1 実験環境

東京農工大学東小金井キャンパス 12 号館交流スペースにおいて 6 分程度の模擬講義を 5 回行ない, これを Cam で撮影した. 被写体となる講師は PC, 液晶プロジェクタとスクリーン, さらに, ホワイトボードも利用した. カメラは図 2 と同じように配置した.

5.2 結果と考察

まず Master カメラの切り替えが生じた時の時間を図 6 に, さらに Sub カメラの切り替えが生じた時の時間を図 7 に示す.

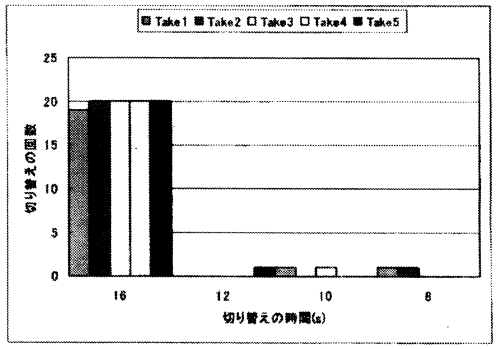


図 6 Master カメラの切り替え

図 6 から Master カメラから Sub カメラへの切り替えは, 16 秒で発生するのが最も多いことが分かる. ショット時間の 16 秒は, このシャットに割り付けられた最大時間であり, 最大時間を越える場合は強制的に他方のカメラに切り替えられる.

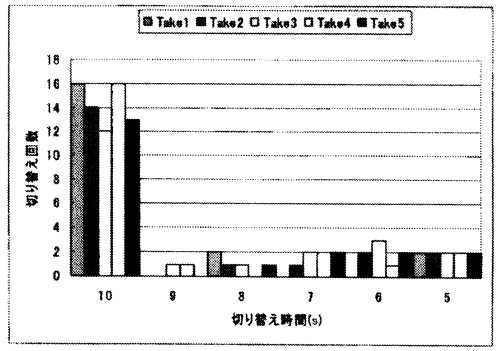
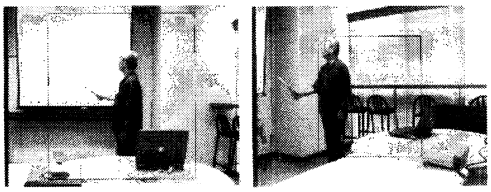


図 7 Sub カメラの切り替え

同じように図 7 から Sub カメラから Master カメラへの切り替えは, 割り付けられた最大時間である 10 秒で発生する頻度が最も高く, 次に最小時間である 5 秒での発生する頻度が続くことが分かる. さらに, 最小時間である 5 秒から最大時間の 10 秒にかけて, 切り替えが発生していることを確認した.

そこで, カメラ間を最小・最大時間以外で切り替えが発生している映像を図 8 に示し, どのような講師の動きを起点として切り替えが生じているのか確認した. a は Master カメラから Sub カメラに 10 秒で切り替わった時, b は Sub カメラから Master カメラへ 6

秒で切り替わった時の映像である．



a　　b

図 8　最小・最大時間以外の切り替え例

a では，講師の全体は画面の中央にあるものの，講師が持つ指示棒の先端部分（映像内の内枠を越えた部分）を Master カメラが追跡を始めたため，Sub カメラへ切り替えられている．同じように b は，指示棒を握る手首周辺部分を Sub カメラが追跡を始めたために，Master カメラへ切り替えられた．これは，sCam の講師認識処理が，フレーム間差分を用いた移動物体の抽出を前提としているため，講師の全身移動と各部位の所作を区別することが出来ないからである．

以上より，映像の切り替えは，割り当てられた最大のショット時間で発生する頻度が最も多く，次に，最小のショット時間で切り替えられる頻度が続く．これは，講師による動きが，カメラの視点を大きく変化させるものよりも，所作・ジェスチャといった微小動作のほうが多いことを示している．

しかし，その微小動作は必ずしも画面中央付近で生じるものだけではなく，図 8 のように中央付近以外で認識した場合，カメラのパン・チルト制御量を変化することにより，結果的に映像切り替えが生じることが分かった．

6. おわりに

本稿では，e-Learning コンテンツとして組み込まれる講師映像を自動的に撮影するシステムを提案，実装した．我々による手法は，従来の講師映像の撮影方法に比べ，容易にカット編集が施され，講師映像として妥当な冷静さと，適度なリズム感を有する映像を作成することが可能となった．使用する機材は，一般的な PC やカメラであるため，比較的低コストでシステムを構築することが可能である．

参考文献

1) 日本イーラーニングコンソシアム編：e ラーニング導入ガイド，東京電機大学出版局，(2004).

2) 西口，東，亀田，角，美濃：講義自動撮影における話者位置推定のための視聴覚情報の統合，電子情報通信学会論文誌 (D-I)，vol. J84-D-I，no. 9，pp. 1421-1430，(2004).

3) 亀田，新，西口，美濃：撮影対象の運動履歴に基づく固定ショット切り替え式撮影法，電子情報通信学会研報 MVE，Vol. IE2003-14，No. MVE2003-44，pp. 1-6，(2003).

4) 大西，村上，福永：状況理解と映像評価に基づく講義の自動撮影，電子情報通信学会論文誌 (D-II)，vol. J85-D-II，no. 4，pp. 594-603，(2002).

5) 日本映画・テレビ編集協会編：図解映像編集の秘訣，玄光社，(1999).

6) 横田：プロフェッショナルビデオ制作技法，映像新聞社，(1990).