データベースシステム 127－7
情 報 学 基 礎 67－7
（2002．5．21）

キーワード毎のショット長分布を用いたビデオ映像シーン検索

吹野直紀† 角谷和俊† 田中克己†

本報告では、区間ではなくある時刻に対して断片的に付いたラベルを基にビデオシーンを検索する手法について述べる。その際、アノテーション時刻とショット区間との関係を表す評価関数のデータを辞書として用意しておき、それを用いて区間推定を行う。複数キーワードによる検索の場合は検出された区間の優先度を決定する際にもその評価関数を用いる。また、検索する際により適した区間を取り出すためのいくつかの工夫について述べる。実装したプロトタイプの評価も行う。

Video Scene Retrieval by Keyword-wise Shot Length Distribution

NAOKI FUKINO ,† KAZUTOSHI SUMIYA † and KATSUMI TANAKA†

In this paper, we propose a method to retrieve a video scene from a fragmentarily-indexed video, in which each index term (called label) is attached not to a video interval but to a time point. We use data concerned with the relationship between an annotated time and a video shot length for each keyword in order to predict the corresponding video interval for each keyword. Furthermore, the data is used to determine the priority of retrieved video intervals. We also describe a way to select pertinent video intervals from a retrieval result. Finally, we describe a preliminary evaluation result of our prototype system.

1. はじめに

今後，ハードディスクを搭載した蓄積型テレビやＤＶＤ録画機等の普及により，個人が所有するランダムアクセス可能なビデオデータの量が増加すると考えられる．また，ネットワークの高速化により動画配信，特に携帯端末への動画配信が実用化されつつある．この様な状況において，膨大なビデオデータから見たい場面を取り出したい場合や，ネットワークから端末へ必要な場面のみダウンロードしたい場合共に必要になるのがビデオのシーンの検索である．しかし，現状ではビデオシーン検索はWWW検索に比べて一般的では無い．

その原因の一つに，問い合わせの難しさを挙げる事が出来る．ビデオデータは非常に多くの情報を含んでいるが故に，ユーザの問い合わせの意図を反映する方法にも様々なものが考えられ，単純に扱う事は出来ない．映っているオブジェクトの動きが重要なビデオに対しては，その動きで問い合わせる方法も研究されている[1]．キーワードで検索する場合も，例えば「人　犬」で検索した場合，犬と人が同時に映っている区間を取り出すべきかもしれないし，ビデオは時系列データであるので，犬と人が連続で映るシーンをユーザは求めているかもしれない．

他の原因には，ビデオの内容記述の困難さがある．元のビデオデータは連続な画像データと音声データのみであり，そこから目的の映像区間を検索するためには何がどの部分にどう映っているかというメタデータを用意する必要がある．そのメタデータは各オブジェクトの位置データであったり，どの映像区間に何が映っているかという索引としてのラベルデータであったりする．メタデータの記述方法は，MPEG7[2]において標準化が進んでいるが，これは記述フォーマットを定めているだけであり，実際にどのような情報をメタデータとして持つのか，また，その検索方法などは定められていない．

このようなメタデータを用意する手法として画像認識や音声認識が役に立つが，人が見ないと理解できない映像もあるので，どうしても人手に頼らざるを得ない部分もある．しかし人手でメタデータを付けるのは大変な労力を要し，そのような内容記述のコストの問題は，特に生中継映像に対してリアルタイムに行う場合に，より深刻である．

本研究ではリアルタイムに内容記述が行われたビデオデータに対してキーワード検索を行う場合を想定し，アノテーション時刻前後のある時点がどの程度の確率で該当シーンであるかという事を表す評価関数（以下，ショット区間評価関数と呼ぶ）と，ラベル間の関連度という２つの概念を用いて「ユーザの入力した全キーワードを高い関連性を持って含む区間」を出力として返す方法を提案し，その有効性について検証する．

以下，２章で基本的事項及び関連研究を述べ，３章では本研究に独特の概念であるショット区間評価関数と関連度についての説明を行う．４章でビデオシーン検索における上記の概念の利用に付いて述べ，５章では実装したプロトタイプについてその評価を行う．ここで，段階的質問変化と関連度の関係についても触れる．６章では結論を述べる．

† 京都大学大学院情報学研究科
Graduate School of Infomatics, Kyoto University

2. 基本的事項

本章では，本報告において用いる基本的事項と関連する研究について述べる．

2.1 ビデオアノテーションとラベル

ビデオデータを検索するためには，そのビデオのある時間に何がどのようにうつっているのかという情報が必要になるので，ビデオの内容を何らかの手段で記述しておく必要がある．これをビデオ映像データのアノテーション (注釈付け) という．ビデオのある時点に対し，それを修飾するキーワードを以下ではラベルと呼ぶ．野球の映像では，ある選手が三振したシーンに対して付いた「三振」というキーワードがラベルとなる．

2.2 構造化法，層状化法

ラベルは，通常ある区間に対して付く．「犬」というラベルは，犬が映っているシーンの初めから終わりまでの区間に対して付くのが理想的である．このような区間の決め方については，大きく分けて2通りの方法がある．ビデオを最初に小さな区間に区切り，それぞれの区間に対してラベルを付けていく構造化法と，ラベルごとにフレーム単位で区間を決める層状化法である．各々のイメージを図で表現すると図1のようになる．構造化法は，映画等ショットやシーンの区切れ目がわかりやすい映像に向いているが，サッカーのような区切れ目の決めにくい映像には向いていない．また，層状化法は構造下方より表現力は高いが，それぞれのラベルに対してフレーム単位で区間を決めなければならないので，その部分で構造化法に比べコストが大きい．

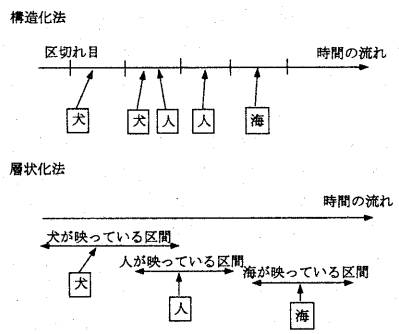


図 1 構造化法と層状化法

2.3 グルー演算

区間同士の演算には，intersection 演算，union 演算等がある．intersection は2つの区間の共通部分を返す．図2では区間 A と区間 B の intersection として区間 R1 を返す事が出来る．union は2つの区間を繋げた区間を返す．図2では区間 A と区間 B の union として区間 R2 を返す事が出来る．しかし，union では2つの区間の間にノイズがあった場合その2区間を繋げる事はしない．図2では，区間 R3 を返す事が出来ない．ビデオにはノイズが入る事はよくあるので，ノイズがあっても2つの区間を繋げて返す glue 演算が提案されている [3]．区間 A と区間 B の glue として区間 R3 を返す事が出来る．区間 A の始めの時間を A_{start}，区間 A の終わりの時間を A_{end} とすると，区間 A と区間 B の glue である区間 R3 は $R3_{start} = min\{A_{start}, B_{start}\}, R3_{end} = max\{A_{end}, B_{end}\}$ となる．以下では，intersection 演算を $\wedge$ で表し，glue 演算を $\bowtie$ で表す．

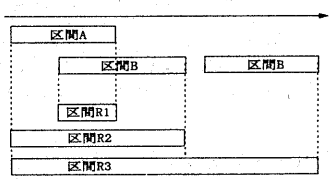


図 2 intersection, union, glue 演算

グルー演算は区間と区間に対する演算だが，これを区間集合と区間集合に拡張したものがペアワイズグルー演算である．2つの区間集合間の任意の区間同士を glue 演算して作った区間の集合を返すものである．図3は，キーワード A に対する区間集合 A1, A2, キーワード B に対する区間集合 B1, B2, キーワード C に対する区間集合 C1 に対して，ペアワイズグルー演算 $A \bowtie B \bowtie C$ を行った結果を表している．

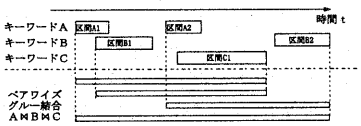
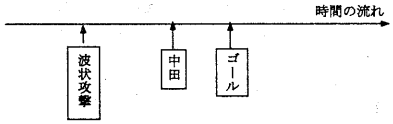


図 3 ペアワイズグルー演算

3. ショット区間評価関数と関連度

ビデオシーン検索の為のラベルの付け方には，区間に対するラベル付けとして大きく分けると構造化法と層状化法があるが，構造化法にはシーンの区切れ目のはっきりしない映像には向かないという欠点があり，層状化法にはアノテーションの際のコストが高いという欠点がある．つまり，例えば生中継サッカー映像に対するアノテーションはどちらの方法でも難しい事になる．このような場合，「区間に対してラベルを付ける」という方法自体を諦めざるを得ない．そこで，以下ではラベルが区間でなくある時点についているという事を前提とする．図4のようなフォーマットでラベルが付いているとする．このラベルは，アナウンサーから音声認識で取ってくる方法や，複数の人手でリアルタイムに付ける方法が考えられる．

このようにラベルに対して区間の情報が無い場合，検索結果として区間を取り出す事が出来ず，AND 検索等も出来ない．これらの問題に対処するため，ショット区間評価



132.45秒　波状攻撃
138.04秒　中田
141.22秒　ゴール

図４　時点に対するラベル付け

関数と関連度という概念を用いる．

3.1　キーワード毎のショット区間評価関数

ショットとはシーンの構成単位であり、通常複数のショットから１つのシーンが構成される。ショットは、ある視点から捉えたカメラ映像であり、サッカーの場合はショットの切り出しそのものが困難であるといえるため、本論文では、サッカーの１つの基本プレイ (例えば、シュートやパスなど) をショットとみなすこととする。このショットの長さは，キーワードによってある程度特徴を持っている．ラベルがアナウンサーの音声によって付いている場合，例えばアナウンサーが「パス」と言った３秒後もまだそのパスが続いている可能性は低い．しかし,「波状攻撃」と言った３秒後もまだその波状攻撃が続いている可能性は高い．図５は，アナウンサーがしゃべった時間を中心とした各キーワードが修飾するシーンが前後どれくらいの区間になっているかを，いくつかのケースについて表している．

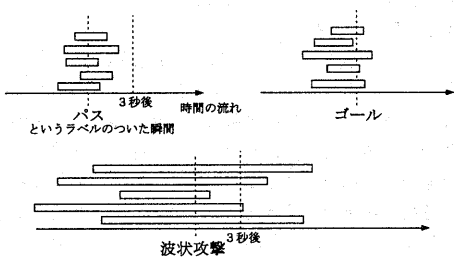


図５　アノテーション時刻を中心としたショットの分布

各キーワード毎の特徴は,「パス」の場合は少し短めになっている.「パス」のショットの長さを仮にパスを出してから受け手が受けるまでだとすると、そのショットの長さは平均して２秒弱であった．また,「波状攻撃」とは元々ある程度長い間攻撃が続いている事を表しているので，この場合は少し長めであり，１０個の波状攻撃シーンをサンプルとして取り出すと，平均して１４秒程度であった．また,「ゴール」の特徴は長さ以外にもある．アナウンサーが「ゴール」と発言するのはゴールを確認した後であり，ゴールのショットの最後の方である．このため,「ゴール」というラベルの付いた時点より前の方にゴールシーンは分布している．これらの特徴を表すために，図６のような評価関数を用いた．横軸はラベルの付いた時刻を原点０とした時間軸であり，縦軸はそれぞれの時刻ｔがラベルに相当するショットに含まれていた確率である．このような評価関数を，本研究ではショット区間評価関数と呼ぶ．ショット区間評価関数のデータは検索の際辞書として持っておく必要がある．☆

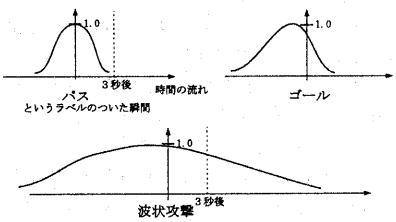


図６　ショット区間の分布を表現する評価関数

3.2　ラベル間の関連度

「中田」というラベルの５秒後に「パス」というラベルが付いていた場合について考える．通常，サッカーでは一人の選手が５秒もボールを持っている事は少なく，５秒後のパスは他の選手によるものである可能性が高い．よって，この「中田」と「パス」の両ラベルは関係がない可能性が高く，関連性は低いと考える．

一方，同じ時間差でも「波状攻撃」というラベルの５秒後に「ゴール」というラベルが付いていた場合，ゴールが波状攻撃の結果によるものである可能性が非常に高い．つまり，この両ラベルの関連性が高いと考えることができる．これは，波状攻撃のショットの粒度が大きい事が関係する．波状攻撃というラベルがついた時間の前後の長い時間が波状攻撃のショットに該当する可能性が高いので，ゴールのラベルがそのショットに含まれる可能性が高くなる．また，同じ「波状攻撃」と「ゴール」というラベルでも，その時間差が３０秒あれば両ラベルの関係は無い可能性が高くなる．

このように，２つのラベルが関係ある可能性については２つの傾向がある．

- ２つのラベルの時間差が同じなら，両ラベルのショットの粒度が大きい方が関係ある可能性が高い．
- ２つのラベルが同じなら，その時間差が小さい方が関係ある可能性が高い．

この両性質を満たすパラメータとして，両ラベルのショット区間評価関数の山が重なる部分で最も高い部分の評価値を使い，これを関連度と呼ぶ事にする．図７ではpである．

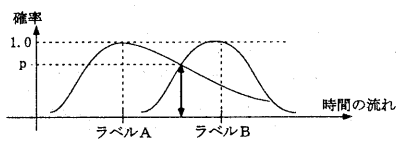


図７　２つのラベルの関連度

☆ この評価関数は，時刻ｔにおける映像がラベルが表すショットに含まれるかどうかを表すメンバーシップ関数に相当すると考えられる

ラベルAとラベルBの関連度がpであるという事実を，以下では

$$Rel\{A,B\}=p$$

と表す．また，関連度の定義を以上のように定義した場合，下の例のような，ラベルの順番の違いによる関係の深さの違いも反映する事ができる．

- コーナーキックの後にゴールというラベルが付いていた場合，そのシーンはコーナーキックがゴールにつながったシーンである可能性が高い．
- ゴールの後にコーナーキックというラベルが付いていた場合，そのシーンはゴールで一度プレーが切れているため，ゴールとコーナーキックには直接の関係は無いと考えられる．

3.3 n個のラベル間の関連度

3つ以上のラベルの関連度について考える場合，基本的に2つの場合と同じく，

- それぞれのラベルが出来るだけ短い時間間隔にまとまっている方が関連度が高い．
- 同じ時間間隔ならそれぞれのラベルに相当するショットの粒度が大きい方が関連度が高い．

という事になる．さらに，例えば「波状攻撃　中田　名波」という検索で中田と名波がからんだ波状攻撃を検索したい場合について考えると，中田と名波に関係が無かったとしても，波状攻撃と中田，波状攻撃と名波に関係があれば，この区間は検索結果として適当である．しかし，波状攻撃と名波，中田と名波が関係ない場合，この区間の中で名波は全く関係が無くなり，このような区間は検索結果として適当ではない．よって，

- 一つでも他の全ラベルと関係の無いラベルを含む区間は関連度を0にする．

という条件も加わる．そこで，ラベルL_1，ラベルL_2，…，ラベルL_nの関連度を以下のように定める．

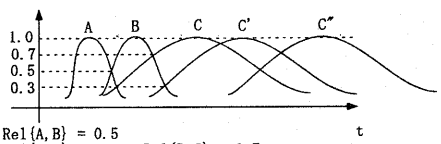
$$\begin{aligned}Rel\{L_1,L_2,...,L_n\}=&\frac{1}{n-1}\sum_{k=1}^{n}Rel\{L_1,L_k\}\\ &\times\frac{1}{n-1}\sum_{k=1}^{n}Rel\{L_2,L_k\}\\ &\times ...\\ &\times\frac{1}{n-1}\sum_{k=1}^{n}Rel\{L_n,L_k\}\end{aligned}\qquad(1)$$

ただし$Rel\{A,A\}=0$である．

キーワードが3つの場合は，下のようになる．

$$\begin{aligned}Rel\{A,B,C\}=&\frac{Rel\{A,B\}+Rel\{A,C\}}{2}\\ &\times\frac{Rel\{B,A\}+Rel\{B,C\}}{2}\\ &\times\frac{Rel\{C,A\}+Rel\{C,B\}}{2}\end{aligned}\qquad(2)$$

つまり，各ラベルに対して他の全ラベルとの関連度の平均を取り，その各々を掛ける．具体例としては，図8のようになる．



Rel{A,B} = 0.5
Rel{A,C} = 0.3,　Rel{B,C} = 0.7
Rel{A,C'}= 0.0,　Rel{B,C'}= 0.3
Rel{A,C''}= 0.0,　Rel{B,C''}= 0.0
Rel{A,B,C} = (0.5+0.3)/2 * (0.7+0.5)/2 * (0.3+0.7)/2
= 0.4 * 0.6 * 0.5
= 0.12
Rel{A,B,C'}= (0.5+0.0)/2 * (0.3+0.5)/2 * (0.0+0.3)/2
= 0.015
Rel{A,B,C''}= 0

図8　n個のラベルの関連度

4. 関連度を用いたシーン検索

本章では，2章で説明したショット区間評価関数，関連度という概念を用いたビデオシーン検索について述べる．4.1節で概要について述べ，以降では検索の際に行う工夫について述べる．4.2節で”AとBは同時に起こらない”という情報等を用いた粒度制御について述べ，4.3節で問い合わせの際のキーワードの順番について，反映させる方法などを説明する．4.4節では，関連度が0になった区間に対して優先順位を付ける方法について述べる．

4.1 概　　要

シーンの検索においては，はじめに質問に相当する区間を取り出し，取り出された複数の区間に優先順位を付ける必要がある．その両方の段階において，ショット区間評価関数を用いる．まず，ビデオ映像から複数キーワードによる検索を行う場合，同じ問い合わせでもユーザによって欲しがる検索結果の区間は異なる．例えば，「波状攻撃　中田　ゴール」という問い合わせをした場合，中田のゴールに繋がった波状攻撃が見たい場合，波状攻撃の途中にある中田のゴールが見たい場合，ゴールに繋がった波状攻撃の途中の中田のプレーが見たい場合，等様々なものが考えられる．これを検索エンジン側で判断する事は出来ないので，全てのキーワードを含む区間を取り出さざるを得ない．全てのキーワードを含む区間というのは何通りも考えられるので，それらに優先順位をつける基準として，本研究では「関連度」とう尺度を定義し，用いる事にする．つまり，検索結果として求められる区間は「全てのキーワードを出来るだけ高い関連性を持って含む区間」になる．

そのような区間を取り出すため，まず始めにショット区間評価関数の評価値が一定値θ以上の部分をそのラベルが表す区間だとする．ユーザが入力した複数のキーワードそれぞれに複数のラベルが一致するので，それぞれのキーワード毎のラベルが指す区間の集合に対して，ペアワイズグルー演算を行う．これで「全てのキーワードを含む区間」の集合をを取り出すことが出来る．キーワードからラベル

の集合を取り出し，それぞれのショット区間評価関数の値 θ 以上の部分をそのラベルが指す区間とし，ペアワイズグルー演算により区間集合を取り出した結果を図 9 に示す．

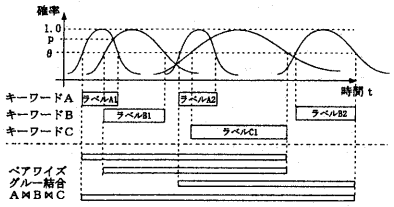


図 9 全てのキーワードを含む区間の取り出し

この算出された区間集合に含まれる区間それぞれに対して，関連度の計算を行う．この関連度をランキング値としてそれぞれの区間に付与し，そのランキング値の高い区間から順に表示する．

4.2 ショット区間評価関数の変形

ラベルには，同時には起こり得ない組み合わせ，つまり，2つのラベルが指す区間が重なる事は有り得ない組み合わせが多く存在する．例えば，サッカーにおいてカウンターとは相手の攻撃をしのいだ後相手の守備が薄いうちに攻める事であるが，この途中に相手が守備位置に帰る事ができるコーナーキックが含まれる事は在りえない．この場合，カウンターのショット区間評価関数を修正してコーナーキック以降の評価値を0にする事で，検索結果から不適切な区間の一部を省く事が出来る．カウンターのショット区間評価関数はすぐ後にある「コーナーキック」ラベルによって図 10 のように変化する．

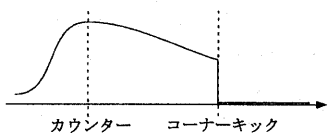


図 10 同時に起こらないラベルによる評価関数の変形

また，ラベルによって単位を決める事ができる場合がある．単位とは，野球で言うとカーブ，空振り等はある1球に対しての修飾である．しかし，三振や新庄はある1打席を修飾するものであるし，三者凡退はある回の攻撃を修飾するものである．そこで，カーブや空振りの単位が1球，三振の単位は1打席，という情報を辞書としてもっておけば，それぞれの単位の切れ目に信号を送る事でショット区間評価関数を修正する事が出来る．

4.3 ラベルの順序

ビデオは時系列データなので，ユーザが「A　B　C」という質問を入力した時，Aの後にBのショットが続き，その後Cのショットが続くシーンを求めているかも知れない．つまり，ユーザが順番を意識している可能性がある．その場合．検索結果にもユーザの意図を反映させたい．

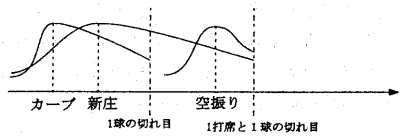


図 11 単位の区切れ目による評価関数の修正

その為に2つのキーワードの関連度 $Rel\{A, B\}$ に，AとBの順番が問い合わせと違っている場合ある値 $w(0 < w < 1)$ を掛ける．これによって，問い合わせとラベルの順序が違っている区間のランキングを下げる事ができる．3つ以上のキーワードに対しても，関連度の計算をする際2つずつのラベルの関連度から計算する事になるので，そのそれぞれについて順番を調べ w を掛ければよい．順番を加味しない場合の関連度に比べ，例を挙げると「A　C　B」という並びの区間は

$$
\begin{aligned}
Rel\{A,C,B\} &= \frac{Rel\{A,C\}+Rel\{A,B\}}{2} \\
&\times \frac{Rel\{A,C\}+w\cdot Rel\{B,C\}}{2} \\
&\times \frac{Rel\{A,B\}+w\cdot Rel\{B,C\}}{2} \\
&= Rel\{A,B,C\} \\
&\times \frac{w^2\cdot {R_{BC}}^2+w\cdot(R_{AB}+R_{AC})\cdot R_{BC}+R_{AB}\cdot R_{BC}}{{R_{BC}}^2+(R_{AB}+R_{AC})\cdot R_{BC}+R_{AB}\cdot R_{AC}}
\end{aligned}
$$

ただし，$Rel\{A,B\}=R_{AB}, Rel\{B,C\}=R_{BC}, Rel\{A,C\}=R_{AC}$

この結果から，Rel{B, C} の値が大きい程，即ち，順番が逆転しているラベルBとラベルCの関連度が高い程本来の順序でラベルが並んだ区間よりも関連度が低くなる事が分かる．

w の値を0にする事で順番が正しくない区間の関連度を大幅に下げ，ランキングを低くする事が出来るが，ラベルが3つ以上の場合関連度が0になるとは限らない．ユーザの意図により順番の正しくない区間を完全に取り除きたい場合は，どこかの2つのラベルの順序が違っている時点で関連度を0に設定する．

4.4 関連度が0の区間の優先順位

関連度が0となる区間の中にユーザが要求している区間が含まれる場合がある．例えば，ユーザが連続で起こる物事に対して検索をかけ，それらの物事に相当するラベルが離れすぎていて関連度が0になっている場合である．そのような区間は通常たくさん検出されるが，関連度が0の区間として同列に扱われてしまう．当然，それぞれのラベルの距離が近い方がよりユーザの要求に見合っているので，関連度が0になっている区間に以下の方法で優先順位を付ける．

(1) 関連度が0ではない区間は，関連度の大きいものから順に出力

(2) それぞれのショット区間評価関数のグラフを時間軸方向に引き伸ばす．

(3) 新たな重なりができる事で関連度が0で無くなる区間が出てくるので1に戻る．

(4) 1～3を繰り返した後，引き伸ばしの倍率がある一定以上になった所で止める．

4は必要以上に処理能力を使うのを避けるためであるが，ラベルの数が少なければ関連度が0になる区間が無くなるまで続ける事も出来る．ショット区間評価関数のグラフを時間軸方向に引き伸ばす操作は，図のようなイメージになる．

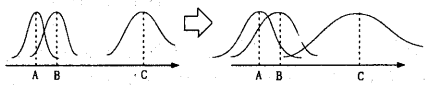


図 12　ショット区間評価関数グラフの引き伸ばし

5. プロトタイプの実装と検証

本研究の有効性を検証するため，プロトタイプの実装を行った．環境を以下に示す．

- OS:Windows2000
- CPU:Pentium4 1.5GHz
- Memory:384MB RDRAM
- 開発環境:Visual C++ 6.0

以下の1節では検索に用いる辞書のデータ構造を示し，2節ではシステムの概要を説明する．3節で検索結果の検証を行い，4節では段階的質問変換との関連について触れる．

5.1 辞書のデータ構造

これまで述べたショット区間評価関数や同時に起こりえないラベル以外に，辞書として持っておくべき情報がいくつかある．

- 同義語
- 単語の上位・下位関係

同義語の情報を持っていなければ，”三浦知良”というラベルに対して”カズ”というキーワードで検索した場合ヒットしない事になる．三浦知良というラベルとカズというラベル両方付けておく事でも解決できるが，リアルタイムアノテーションにおいてラベル付けの冗長性は極力省くべきである．

また，単語の上位・下位関係も必要である．例えば，ボレーシュートはシュートの一種であるが，ボレーシュートというラベルが付いているシーンに対してシュートで検索した場合ここはヒットしなければならない．このような同義語・上位下位関係を持った辞書の例としては，オンライン英英辞書である WordNet[4] がある．

以上の事を踏まえて，辞書データは単語の同義語・上位下位関係を記述したシソーラスにショット区間評価関数データを付加する方法で作った．図13のような構造をとった．

同義語は1つのノードにまとめ，それぞれ上位のノードへのポインタを張っている．ショット区間評価関数のグラフが定義されていないノードは，上位のノードを辿ってショット区間評価関数のデータを得る．上位のノードど辿ってもショット区間評価関数のデータが無い場合は，デフォルトのグラフを用いる．

5.2 システムの概要

システムの処理のおおまかな流れを説明する．

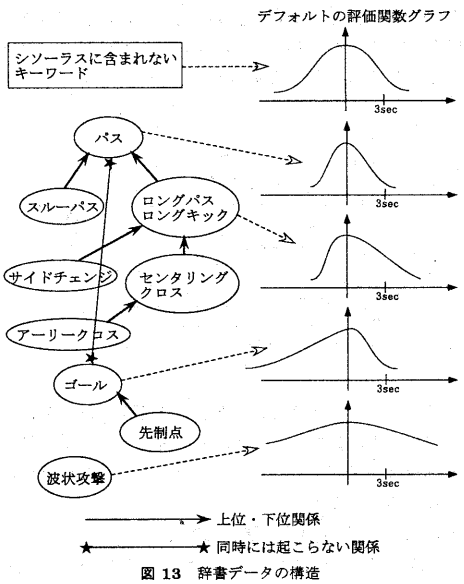


図 13　辞書データの構造

(1) まず，問い合わせとして入力されたキーワードとアノテーションデータ，及び単語の同義語・上位下位関係のデータから，問い合わせに一致するラベルの組み合わせを全て取り出す．

(2) それぞれのラベルの組み合わせに対し，ショット区間評価関数を参照して関連度を計算し，その関連度によって順序を並べ替える．

(3) ユーザが順位を入力した後表示ボタンを押せば，その順位にあるラベルの組み合わせに対し，まず各ラベルが指している区間をショット区間評価関数により計算し，ラベルが指す各々の区間をグルー演算した結果の区間を表示する．

5.3 検　　証

図15のようなGUIのプログラムを作成した．ソフトを起動後，まずビデオファイル，辞書ファイル，アノテーションファイルを読み込み，検索欄に複数のキーワードをスペースで区切る形で入力して検索をかける．ヒット数が表示されるので，右下のボックスにランキングの順位を入力して表示ボタンを押せば入力した順位の区間が表示される．

このプログラムの他にアノテーションファイルと辞書ファイルを用意する必要があったが，アノテーションは時刻とラベルを記述するだけでよかったので，そのためのツールを作ってアノテーションを行った所，1時間で５１０個のラベルを付ける事が出来た．ほぼ7秒に1つのペースであり，生中継映像に対して付けるべきラベルは，どの程度詳しく行うかにもよるが，1秒に1つ程度付けることが出来れば十分であると考えると，7人程度のグループによりリアルタイムアノテーションが実現できる可能性がある．また，アノテーション用に作ったツールは非常にシンプル

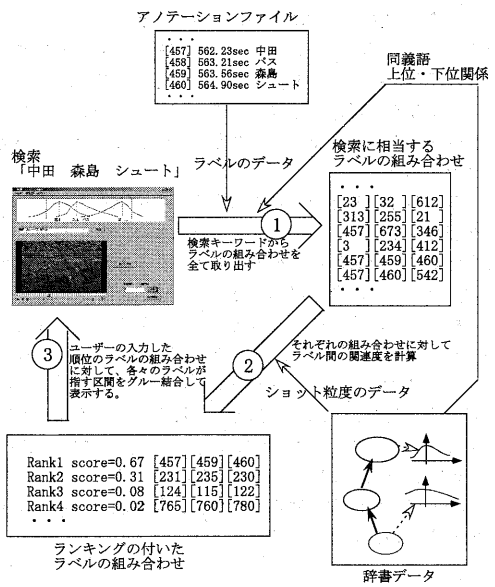


図 14 処理の流れ

であり，全てキーボードで打ち込む方式だったので，キーボードの各ボタンによく使うラベルを割り当てるなどの工夫によりもう少し効率を上げる事はできると考えられる．

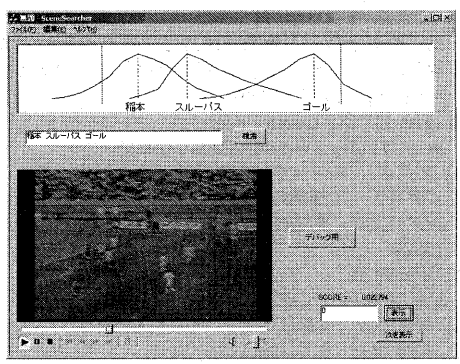


図 15 プロトタイプ

５１０個のラベルを持ったアノテーションファイルに対する検索では，問い合わせとして入力したキーワードの個数に関係なく，検索にかかった時間は体感できる程は無く，検索ボタンを押すと同時に検索は終わるといった感覚だった．サッカーの１試合９０分に１秒１個のペースで細かくラベルを付けると，５０００個以上のラベルが付く事になるが，内部のアルゴリズムでは処理時間はラベル数に対して指数関数的に処理量が増加するわけではなく，線形に近い処理量の増え方であるはずなので，この検索方式において検索時間が問題になる事はあまり無いと考える．

検索結果を見ていると，ほぼ上位のものほど問い合わせの意図に沿ったものが出てきており，下位になるほど問い合わせの意図とは違ったものが出て来ていた．しかし，ほとんど同じシーンが複数ヒットしている場所もあった．

5.4 段階的質問変換と関連度の関係

ラベルが指す区間に対して，intersection 演算や glue 演算によって問い合わせを行う方法がある．ここで，区間 A と区間 B の intersection とは区間 A と区間 B の重なっている部分であり，「A ∧ B」で表し，glue とは区間 A と区間 B を間にノイズがあったとしてもつなげた区間であり，「A ⋈ B」で表す．

ユーザがビデオ映像に対して複数キーワードで検索した場合，そのキーワードが同時に起こっている事，即ちそのキーワードが指す区間が全て重なっている部分がある事を要求しているとは限らない．例を上げると，「中田　ゴール」と検索した場合中田のゴールを検索していると捉えるのが一般的だが，中田のスルーパスからゴールが生まれた場合等，中田が絡んだゴールという意味の検索かもしれない．中田のゴールを探すため「中田 ∧ ゴール」で算出される区間を探しそれが無ければ中田が絡んだゴールという事で「中田 ⋈ ゴール」と条件を緩めて検索する方法がある．

キーワードが２つではなく３つ以上の場合は，まず「A ∧ B ∧ C ∧ D」と全てが重なっている部分を探し出し，無ければ「(A ⋈ B) ∧ C ∧ D」というようにどれか２つのキーワードに対しては glue を取るような質問に変え，それでも無ければ３つ以上の単語の glue 演算を行い，といったように，始めの厳しい質問で検索結果が無ければ少しずつ質問を緩めて検索していく手法を，段階的質問変換と呼ぶ．この手法と関連度による検索手法を比較するため，「中村　センタリング　シュート」というキーワードで検索を行ってみた所，上位５件は図 16 のようになった．それぞれの場合のシーン粒度グラフの重なりは出力画面に表示されるので，５件の出力画面を付録に付けた．

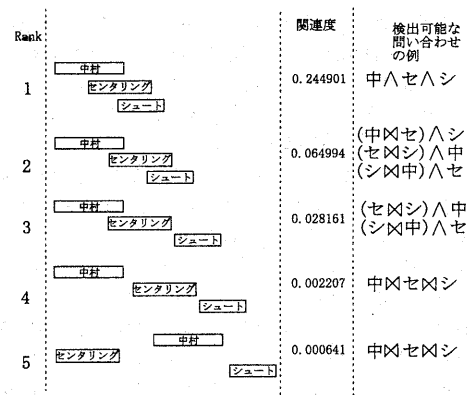


図 16 検索結果の関連度と検出可能な問い合わせ

１件目及び２件目は質問の意図通り，中村選手のセンタリングを他の選手がシュートしたシーンだった．３件目は

中村選手のセンタリングを他の選手がヘディングでまた別の選手にパスし，その選手がシュートしたシーンであり，センタリングとシュートの間に１クッション入った．４件目は中村選手のセンタリングを受けた選手がしばらくドリブルした後，シュートしたシーンであった．最後の５件目は，他の選手のセンタリングの後，中村選手がスルーパスをし，それを受けた選手がシュートしたシーンであり，センタリングとシュートは直接関係は無かった．

これを $\wedge$ や $\bowtie$ で検索した場合を考える．それぞれのラベルが指す区間を”ショット区間評価関数の値が $\theta = 0.4$ 以上の区間”と決めた時の各区間の重なり方は図 16 に描いた．

１件目では全てが重なっている部分があり，「中村 $\wedge$ センタリング $\wedge$ シュート」で検出する事が出来る区間である．２件目は「(中村 $\bowtie$ センタリング)$\wedge$ シュート」のように，任意の２単語をグルー結合する事でヒットさせる事が出来る．３件目も２単語のグルー結合により検出する事が出来るが，任意の２単語というわけではなく，「(中村 $\bowtie$ センタリング)$\wedge$ シュート」という質問では検出する事が出来ない．これは，２件目と違いセンタリングとシュートが離れている事が原因である．４件目以降は，全区間を glue 結合しなければ検出できない区間であった．

以上の事から，関連度の高い区間から順位見ていくと，段階的質問変換をしていく過程と似ている事が分かった．言い換えれば，段階的質問変換により出てくる区間の順序と，関連度の高い順はほぼ一致すると考えられる．関連度の順序と質問変換の段階の順序が逆転するケースも考えられるが、あくまで特殊なケースとなる。これは，関連度が任意の２単語の指す区間が離れる事により大幅に下がる事が原因である．

6. おわりに

本報告では，リアルタイムアノテーションによって区間では無く時間に対してラベル付けされたビデオを対象に，ショット区間評価関数及び関連度の概念を用いて検索を行う方法を提案した．実装では，サッカー映像に対して低コストのアノテーションと正確な検索結果を両立する事ができ，生中継サッカー映像のようにリアルタイムにアノテーションを行う必要がありかつショットごとの区切れ目が曖昧な映像に対して本研究の手法が有効である事を示せた．しかし，いくつかの難しい問題が残った．

まず，低コストなアノテーションとリアルタイムアノテーションにはまだ差がある．リアルタイムにアノテーションを行う場合，どうしても複数人が同時に行う必要が出る．本研究ではアノテーションのタイミングのずれについては考慮しているものの，アノテーションのラベルに間違いは無く，重複もないという前提に立っている．しかし，複数人によってアノテーションを行う場合どうしても重複による冗長性や，間違いによる矛盾が生ずる．このような冗長性及び矛盾をショット区間評価関数情報付きシソーラスを用いる事で解決する方法について検討したい．

また，例えばパスのシーンは多くの人がパスを出してから受け手が受けるまでをパスシーンだと考えるが，ゴールシーンになるとどこからどこまでをゴールシーンと解釈するかは人によって異なる．この点の曖昧さについても考える必要がある．

その他の今後の課題としては複数の生中継映像のザッピング視聴が挙げられる．サッカーでは公平を期する為等の理由により，複数の試合が同時に行われる事が多い．現状では，どれかの試合を選んで見ながら適当にチャンネルを切り替えるか，途中経過の報告によって他の試合の展開を知るかしか無い．しかし，もしそれぞれの試合がリアルタイムにアノテーションされていたとしたら，何らかの方法でユーザの興味を記述しておき，それをもとに検索をかけながらランキング値の高い区間を連続的に流していく事で，多くの試合を同時に楽しむことが出来る．そこで，ユーザの興味の記述の仕方，検索のかけ方，複数映像のバッファの取り方等の検討を行う予定である．

謝　辞

本研究の一部は，文部科学省科学研究費基盤 (B)(2)「蓄積型放送のためのパーソナル視聴の研究」(課題番号 14380177) による．ここに記して謝意を表します．

参考文献

1) 矢島史, 中西吉洋, 田中克己：
動きの直接指定と時間関連指定に基づく移動体映像検索，情報処理学会 DBS 研究会技術報告, Vol.2001, No.71 pp.169-176, 2001
2) The MPEG Home Page:
http://mpeg.telecomitalialab.com/
3) プラダン スジット, 田島 敬史, 田中 克己：
ビデオデータ検索のための区間グルー操作と解のフィルタリング，情報処理学会論文誌・データベース (Jan, 1999)
4) WordNet:
http://www.cogsci.princeton.edu/ wn/
5) 岩波書店　情報の構造化と検索 (p2〜p23)
6) 吹野 直紀, 角谷 和俊, 田中 克己:
アノテーション時刻とショット長の確率分布に基づく粒度可変型ビデオシーン検索, 第６４回情報処理学会全国大会論文集第３分冊, 215-216, 2002